VNIVERSITAT DE VALÈNCIA

FACULTAD DE DERECHO

Departamento de Derecho Administrativo y Procesal

Programa de Doctorado en Derecho, Ciencia Política y Criminología

TESIS DOCTORAL

Posibilidades actuales y futuras para la regulación de la discriminación producida por algoritmos

Presentada por:

Alba Soriano Arnanz

Dirigida por:

Prof. Dr. Andrés Boix Palop

Profesor Titular de Derecho Adminsitrativo

València, octubre de 2020

INTRODUCCIÓN

1. JUSTIFICACIÓN Y DELIMITACIÓN DEL OBJETO DE ESTUDIO

El tema objeto de estudio de esta tesis doctoral es el análisis de las posibilidades actuales y futuras para la regulación de la discriminación algorítmica.

Un algoritmo constituye una serie de instrucciones dirigidas a la resolución de un problema paso a paso. En el contexto de esta investigación, los algoritmos estudiados se ejecutan por ordenadores. Por ejemplo, un algoritmo contenido en el sistema informático de una universidad puede servir para contar las personas que se matriculan en un curso.[1] El algoritmo incorporará las siguientes instrucciones: teniendo en cuenta que, inicialmente, n=0, realizar la operación n + 1 cada vez que aparezca un nuevo nombre en la lista, convirtiendo, en cada ocasión, el resultado de esa operación en el nuevo valor de n. Es decir:

Inicialmente, N=0

Por cada nuevo nombre en la lista realizar la siguiente operación:

N+1=Nuevo valor de N

Estas instrucciones se traducen y operacionalizan en código (lenguaje informático). Los algoritmos se utilizan para la consecución de una finalidad en el marco de una representación de la realidad. Esa representación de la realidad, construida mediante las instrucciones contenidas en los algoritmos, se denomina modelo. Se construye un modelo empleando un algoritmo (serie de instrucciones) para contar personas matriculadas. Tenemos un modelo informático (una representación) de lo que sería una persona humana contando alumnas y alumnos matriculados.

El algoritmo arriba explicado es muy sencillo. Sin embargo, el desarrollo tecnológico y la creciente capacidad computacional hace que los sistemas automatizados sean cada vez más complejos. En la actualidad, estos sistemas pueden rápidamente analizar enormes cantidades de datos, inasumibles para cualquier persona humana, e incluso mejorar la forma en que realizan estos tratamientos de datos, en lo que es un proceso claramente semejante al

[1] Este ejemplo se encuentra inspirado en las explicaciones de los conceptos de modelo y algoritmo dadas por David Malan y Cathy O'Neil en: MALAN, D., "What is an algorithm?", Mayo de 2013. Disponible el 27 de

aprendizaje, de manera autónoma. Es por ello que estos sistemas se emplean de manera creciente en toda clase de procesos de toma de decisiones con el objetivo de clasificar o predecir situaciones o actuaciones de personas, recomendar líneas de actuación, o una combinación de todas estas funciones.

Los algoritmos se suelen ejecutar agrupados en programas. Es por ello que las referencias a algoritmos realizadas a lo largo de la tesis deben entenderse como referencias tanto a algoritmos individualmente considerados como agrupados en programas.

El auge de los sistemas automatizados de toma de decisiones en todos los ámbitos ha traído consigo la constatación de que, lejos de resolver problemas y procesos de manera objetiva, los algoritmos reproducen y perpetúan las estructuras de discriminación que afectan a las personas pertenecientes a grupos que, históricamente, se han encontrado en una posición de subordinación y desventaja. Esta tesis doctoral se centra en las denominadas categorías sospechosas de producir discriminación, es decir, aquellas características que se corresponden con elementos de la persona en principio inmutables, como el sexo, la edad o la raza, o que se sitúan en el núcleo mismo de la dignidad de la persona, como las creencias religiosas u opiniones políticas. Precisamente por ser estas características inmutables o íntimamente relacionadas con la dignidad de la persona se atribuye un mayor desvalor a las decisiones que discriminan con base en dichas características de la persona.

Dentro de las categorías sospechosas, la presente tesis se centra en la protección de los grupos especialmente desaventajados o vulnerables, a saber, personas de bajo nivel socioeconómico; mujeres; minorías étnicas, raciales, nacionales y religiosas; grupos de edad vulnerables; identidades y orientaciones sexuales no normativas; personas con discapacidad, etc. Es decir, en las personas pertenecientes a aquellos grupos que han sufrido una situación histórica de desventaja que todavía hoy se perpetúa en la construcción de las estructuras sociales de poder.

Los algoritmos y las nuevas tecnologías de procesamiento de datos generan también una amplia variedad de problemas que van más allá de las posibles vulneraciones a los derechos a la igualdad y no discriminación con base en las llamadas “categorías sospechosas”. Como se verá a lo largo de la tesis doctoral, los sistemas automatizados presentan problemas de opacidad, de dificultad en la adjudicación de la responsabilidad de las decisiones, así como importantes riesgos para la autonomía, libertad y dignidad de las personas y sus derechos a la

intimidad y a la protección de datos. Asimismo, cuando estos sistemas se emplean por el sector público, pueden conllevar importantes quiebras en la cadena de legitimidad de las decisiones emanadas de los poderes públicos.

Todos los problemas derivados del creciente uso de algoritmos se encuentran, además, íntimamente ligados entre sí. A mayor abundamiento, en la actualidad, el marco normativo en materia de protección de datos constituye la principal herramienta dirigida a regular, de manera específica, el tratamiento y procesamiento automatizado de datos. Las normas de protección de datos pretenden ofrecer un amplio marco que, en la medida posible, abarque todas las vicisitudes derivadas de la progresiva automatización de procesos. Es por ello que, si bien las cuestiones problemáticas derivadas del uso de sistemas automatizados citadas en el párrafo anterior no constituyen, en sentido estricto, el objeto de estudio de esta tesis doctoral, también lo son y se abordan, en la medida en que afectan a las situaciones de discriminación y perpetuación de la desigualdad mediadas por sistemas automatizados.

Cabe también destacar que el objeto de estudio de la tesis no lo constituye cualquier clase de tratamiento o de procesamiento de datos, sino que solo lo constituyen aquellos tratamientos y procesamientos de datos realizados de manera total y parcialmente automatizada. Es importante delimitar esta cuestión por cuanto la normativa en materia de protección de datos, que se estudiará en la segunda parte de la tesis, también aborda el tratamiento no automatizado de datos. Por ello, se debe tener en cuenta que toda referencia realizada a las disposiciones normativas contenidas en dicho marco jurídico y a los problemas derivados del tratamiento y procesamiento de datos se debe entender efectuada al tratamiento, procesamiento y consiguiente toma de decisiones automatizadas.

2. METODOLOGÍA Y ESTRUCTURA

2.1. METODOLOGÍA GENERAL Y RAMAS JURÍDICAS EN LAS QUE SE ENMARCA LA TESIS

La presente tesis doctoral se basa en las líneas metodológicas propias de la investigación jurídica. La investigación jurídica, como disciplina, se dirige al análisis y descripción de las normas jurídicas y sus efectos sobre la realidad social partiendo de la recopilación y análisis de bibliografía jurídica, jurisprudencia y textos normativos. Este trabajo se enmarca en el ámbito del Derecho público, situándose en los puntos de conexión entre las ramas del Derecho Administrativo y el Derecho Constitucional.

Ahora bien, en esta tesis doctoral se parte de la noción del ordenamiento jurídico como institución social que debe evolucionar y aproximarse a la realidad que regula. Es por ello que se realiza una aproximación al fenómeno de la discriminación de los grupos desaventajados, como base de la discriminación algorítmica, desde un marco teórico basado en las teorías críticas de la igualdad, eminentemente construido desde la perspectiva de la teoría y filosofía jurídica y política.

2.2. MARCO DE ANÁLISIS SUPRANACIONAL

Asimismo, el marco jurídico de análisis de este trabajo se centra en el ordenamiento de la Unión Europea (y también la Convención Europea de Derechos Humanos como instrumento jurídico de protección de los derechos a la igualdad, a la no discriminación y a la protección de datos, aplicable a los países miembros de la Unión Europea).

La razón por la que se opta por un ámbito de análisis supranacional es, en primer lugar, que el fenómeno de la extensión del uso de sistemas automatizados, y los problemas que dichos sistemas generan, tiene lugar de manera muy similar en todos los Estados europeos.

En segundo lugar, el ámbito de acción de los sistemas algorítmicos no se rige por las fronteras nacionales, sino que un mismo sistema puede afectar a ciudadanas y ciudadanos de distintos Estados. Además, no debe perderse de vista, reforzando esta idea, que un sistema creado en un Estado puede ser vendido a empresas pertenecientes a otros muchos, lo que abunda en la necesidad de un tratamiento y encuadre jurídico que vaya más allá del ámbito estrictamente nacional. En definitiva, y como es sabido, en el caso de las grandes corporaciones tecnológicas transnacionales, cada vez más presentes y dominantes, los problemas generados por los algoritmos que estas empresas emplean tienen incidencia a nivel global.

En tercer lugar, y por último, las respuestas normativas que se ha pretendido dar a los problemas generados por esta clase de sistemas y que pueden englobarse en la elaboración del marco jurídico en materia de protección de datos, han tenido lugar a nivel europeo. Además, las recientes propuestas relativas a un nuevo marco de regulación y control de los algoritmos también prevén que las futuras respuestas jurídicas a los riesgos y daños generados por los algoritmos se desarrollen por las instituciones de la Unión.

Cabe también destacar que, con el objetivo de dotar de una vertiente lo más práctica posible al presente trabajo, se extraen numerosos ejemplos de sistemas algorítmicos empleados no solo en los Estados de la Unión Europea, sino también en otros países, sobre todo en Estados Unidos. Es importante tener en cuenta que los problemas derivados de los sistemas automatizados empleados en EE.UU. no constituyen el objeto de estudio de esta tesis doctoral. Sin embargo, es de gran relevancia hacer referencia a ellos, e incluso determinar la forma en la que el ordenamiento jurídico europeo de protección a los derechos a la igualdad y a la no discriminación podría dar respuesta a los casos de discriminación algorítmica surgidos en ese país o provocados por empresas o individuos que actúen dentro del marco de acción del Derecho estadounidense. Esto es así por cuanto las innovaciones tecnológicas, y los cambios sociales que conllevan, que tienen lugar en Estados Unidos, tienden a expandirse al resto del mundo y, en el caso del creciente uso de sistemas automatizados tanto por el sector público como por el sector privado, esta realidad ha sido especialmente patente.

2.3. HIPÓTESIS GENERALES DE TRABAJO

La primera hipótesis de la que se parte al abordar la realización de la presente tesis doctoral es que las normas jurídicas existentes que pueden ser empleadas en la protección frente a la discriminación algorítmica y la perpetuación de la desigualdad producida por sistemas automatizados no ofrecen a la ciudadanía, en la actualidad, una garantía de protección suficientemente satisfactoria.

La segunda hipótesis sobre la que se basa este trabajo es que sí existen mecanismos jurídicos que permitirían desarrollar mecanismos más efectivos de protección frente a la discriminación algorítmica y la perpetuación de la desigualdad producida por sistemas automatizados, así como que la adopción de los mismos ha de ser contemplada a corto plazo por nuestras sociedades y poderes públicos, especialmente dada la magnitud de los riesgos que pueden suponer, especialmente para los colectivos más vulnerables. Esta tesis doctoral aspira a aportar elementos de análisis jurídico que permitan avanzar, precisamente, en esta tarea.

2.4. NORMAS JURÍDICAS ANALIZADAS Y ESTRUCTURA DE LA TESIS DOCTORAL

La discriminación algorítmica es una realidad que debe ser analizada desde una doble vertiente o perspectiva. Por una parte, como fenómeno que vulnera el principio de igualdad y

los derechos fundamentales a la igualdad y a la no discriminación y, por otra, como uno de los múltiples daños y riesgos generados por las tecnologías de procesamiento de datos y de automatización de la toma de decisiones. La tesis se divide en dos partes que, respectivamente, se ocupan de cada una de estas perspectivas de análisis.

Con respecto a la primera perspectiva, resulta necesario entender la construcción de las sociedades actuales sobre unas estructuras que, históricamente, han situado a las personas pertenecientes a determinados grupos en una posición de desventaja. Dichas estructuras de discriminación generan los fenómenos de discriminación estructural, sistémica o institucional que todavía hoy perviven y que sitúan a determinadas personas en una posición de desventaja y especial vulnerabilidad como consecuencia de su pertenencia a determinados grupos. La construcción de unas sociedades sesgadas (discriminatorias) tiene como resultado la creación de instituciones y productos, también los normativos, desequilibrados a favor de los grupos dominantes y su ideología. Los algoritmos, como productos creados por los seres humanos y utilizados para analizar, reflejar la realidad y tomar decisiones con respecto a la realidad que analizan, corren el riesgo de integrar e interiorizar, por una amplia serie de razones, esos mismos sesgos y estereotipos que perjudican a las personas pertenecientes a grupos desaventajados. Como forma de vulneración de los derechos a la igualdad y a la no discriminación, la discriminación algorítmica debe ser, por tanto, abordada desde el marco jurídico en materia de igualdad y no discriminación.

La primera parte de la tesis se divide en cuatro capítulos fundamentalmente dirigidos a explicar el fenómeno de la discriminación algorítmica desde la perspectiva del ordenamiento jurídico en materia de igualdad y no discriminación y las teorías críticas de la igualdad. Asimismo, se ha considerado conveniente introducir un primer capítulo que, de manera breve y superficial, expone y explica diferentes conceptos relevantes relativos a las tecnologías analizadas en esta tesis doctoral, demuestra su creciente utilización tanto en el sector público como en el sector privado y analiza algunos de sus posibles usos y aplicaciones así como sus riesgos. En este primer capítulo también se abordan, de manera general, los diferentes riesgos y daños derivados de estos sistemas, remitiéndose a la segunda parte de la tesis para su mayor desarrollo. Por su parte, el segundo capítulo de esta primera parte explica la construcción de la discriminación como fenómeno social y jurídico. A continuación, el tercer capítulo analiza la forma en la que se produce la discriminación algorítmica y, finalmente, el cuarto y último capítulo examina las posibles soluciones que pueden darse frente a las situaciones de

discriminación algorítmica desde el actual marco jurídico europeo en materia de protección de los derechos a la igualdad y no discriminación, así como las deficiencias de dichas normas. Algunas de las carencias analizadas son comunes a todas las formas de discriminación y otras surgen específicamente a raíz de las particularidades de la discriminación mediada por algoritmos.

Con respecto a la segunda vertiente desde la que se debe analizar la discriminación algorítmica como fenómeno social y jurídico, cabe destacar, como ya se ha indicado, la íntima relación que guardan entre sí los diferentes riesgos y afecciones a los derechos fundamentales y otros intereses públicos derivados del creciente uso de algoritmos. Así, por ejemplo, difícilmente podrán probarse los casos de discriminación algorítmica directa, esto es, la toma de una decisión discriminatoria teniendo en cuenta una categoría sospechosa, si no se tiene acceso o no se dispone de una explicación suficiente de la lógica sobre la que basa su decisión el sistema. Así pues, considerando la vocación del marco jurídico en materia de protección de datos de hacer frente de manera directa o indirecta a los diferentes daños y riesgos generados por los sistemas de procesamiento de datos y de toma de decisiones automatizadas, resulta necesario dedicar una parte de este trabajo al análisis de dicho instrumento jurídico.

Desarrollando esta idea, la segunda parte de la tesis se ocupa, en su primer capítulo, de analizar los diferentes riesgos y daños derivados de los sistemas algorítmicos de toma de decisiones. El primer capítulo de la segunda parte también establece y explica las razones por las que, hasta la fecha, se ha optado por el marco jurídico en materia de protección de datos como principal herramienta para hacer frente a parte de los problemas surgidos del creciente uso de algoritmos. El segundo capítulo analiza el marco jurídico europeo en materia de protección de datos. El tercer capítulo establece las limitaciones e insuficiencias de dicho marco jurídico a la hora de hacer frente a la discriminación algorítmica y a otros de los problemas generados por el creciente uso de sistemas automatizados. Finalmente, el último capítulo de esta segunda parte de la tesis desarrolla, de manera breve, una serie de propuestas, basadas en técnicas jurídicas disponibles y normas jurídicas ya existentes, dirigidas a lograr una mejor regulación y control de la discriminación algorítmica. La razón por la que este último capítulo se ubica en la segunda parte es que, a pesar de que su objetivo fundamental es proponer un sistema para la prevención y protección frente a la discriminación algorítmica, el sistema propuesto también aspira a servir como base para una regulación y control general

frente a todos los riesgos para los derechos fundamentales y los intereses públicos derivados del uso de sistemas automatizados.

PART I. ALGORITHMIC DISCRIMINATION

Algorithmic discrimination, as a double-pronged phenomenon, must firstly be addressed from the perspective of the theoretical and legal frameworks of protection to the rights to equality and non-discrimination in order to conceptually define what the object of study means and constitutes from a legal perspective. However, before delving into the different elements that define the equality and non-discrimination framework and its applicability in cases of algorithmic discrimination, it is necessary to establish a general overview of what algorithms are and how they are being employed.

Hence, the first chapter offers a brief outline of the way in which algorithms work, some of the applications and purposes for which they can be used and the harms and shortly addresses risks (probability of causing harm)[2] they can generate. The more comprehensive explanation of the concerns, harms and risks generated by automated decision-making, other than discrimination, is carried out in the first chapter in part two due to the fact that the second part of the dissertation analyses the regulatory instruments that aim to deal, to a larger o lesser extent, with all the risks and harms generated by algorithms whereas this part specifically focuses on the equality and non-discrimination legal protection framework.

The second chapter sets out the theoretical framework for the legal protection of the rights to equality and non-discrimination and explains how structures of discrimination that have historically placed the members of certain groups at a position of disadvantage have been constructed. The third chapter explains some the processes that can lead to algorithms yielding discriminatory results and perpetuating structures of inequality. The fourth chapter addresses the way in which algorithmic discrimination can be addressed from the perspective of the European equality and non-discrimination framework and the fifth and final chapter analyses some of the limits and shortcomings of said framework.

[2] DOMÉNECH PASCUAL, G., *Derechos Fundamentales y Riesgos Tecnológicos*, Madrid, Centro de Estudios Políticos y Constitucionales, 2006, pp. 250-251.

CHAPTER I. AN INTRODUCTION TO ALGORITHMIC DECISION-MAKING

Automated data processing and decision-making systems have existed and been used by both the public and private sector for decades.[3] Earlier systems, some of which are still useful for a number of purposes, worked by matching different databases, each of which contained a very specific set of data, such as information on taxes or social security. For example, in Norway, the housing aid system has been automatised since 1972 in order to make aid assignment more efficient[4] and in the Netherlands, a system that matches two simple databases in order to fine drivers has been put in place for quite some time.[5] In addition, simple forms of automated recruitment systems were also developed and deployed since the 1970s.[6]

However, the current capacity of the technologies involved in these systems for processing and generating information,[7] that is, the computational power that is now available, has largely extended the use of these tools. Automated systems are used in a wide variety of contexts, from predicting recidivism risk to targeting ads to certain groups or individuals.[8] Their heavy penetration in all areas of Western societies has increased the number of ways in which they can directly and indirectly affect individuals' fundamental rights as well as many aspects of their lives. In addition, these systems are becoming more and more complex and, sometimes, almost impossible to understand and control by humans.[9]

All of these issues, their increasing capacity, widespread use, the risks they generate, their lack of transparency and difficulties regarding their control, bring one of the key questions to which this dissertation aims to answer: are existing regulatory instruments appropriate to address the problems generated by these newly developed technologies, or should new regulations that specifically address algorithmic decision-making be developed?

---

[3] PADOFF, R., "Why deep learning is suddenly changing your life", *Fortune*, 28th September 2016. Available on 28th April 2019 at: http://fortune.com/

[4] BING, J., "Code, access and control" in MURRAY, A., & KLANG, M., *Human Rights in the Digital Age*, London, Glasshouse Press, 2005, p. 204.

[5] VAN ECK, M., "Algorithms in public administration", 31st January 2017. Available on 17th July 2019 at: https://marliesvaneck.wordpress.com/

[6] LOWRY, S. & MACPHERSON, G., "A blot on the profession", *British Medical Journal*, 5th March 1988, pp. 657-658.

[7] PADOFF, R., "Why deep learning is suddenly changing your life", *cit.*, 2019.

[8] O'NEIL, C., *Weapons of Math Destruction…*, *cit.*, 2017.

[9] BHATIA, R., "How do machine learning algorithms differ from traditional algorithms?", *Analytics India Magazine*, 10th September 2018. Available on 13th June 2019 at: https://analyticsindiamag.com/

This chapter offers a general overview of the technologies used in automated data processing and decision-making with regard to groups and individuals. A brief summary of some applications in which algorithms are used is also carried out with the objective of showing the extent to which these systems have penetrated our societies and affect individuals. The final part of the chapter is dedicated to discussing the different types of algorithms, the purposes they can serve and also very briefly and generally addresses some of the problems and risks generated by these systems. The purpose of this chapter is thus simply set the scene on which the thesis is built on, by explaining the technologies that are the object of this study.

1. BIG DATA

Before delving into the issues that arise from the use of automated decision-making and putting forward proposals regarding the ways in which said problems could be addressed, it is necessary to briefly analyse and explain the current state of technology. One of the key elements that underlie the penetration of automated systems in many areas of society is the elusive concept of "big data". The rapid technological developments undergone over the past two decades have led to the widespread use of the phrase "big data" in a very short period of time over all kinds of disciplines[10] and areas of society.[11] However, use of the concept has extended without a consistent and single definition being attributed to this relatively new phenomenon.[12]

Nonetheless, there are three elements that can be drawn from most comprehensive definitions of big data which help understand not only what big data is but also its origin and the implications it has for society:

- The large volume, velocity and variety of the data (moving forward referred to as 'the three Vs');
- The need to process this type of data; and,
- The value of the results big data provides once it has been processed.

[10] DE MAURO, A. *et al.*, "What is big data? A consensual definition and a review of key research topics", paper presented at the *4th International Conference on Integrated Information*, Madrid, 5-8th September 2014, p. 97.
[11] WARD, J., & BARKER, A., "Undefined by data: a survey of big data definitions", 2013, p. 1. Available on 20th September 2018 at: https://arxiv.org/
[12] FRANKS, B., *Taming the Big Data Tidal Wave,* Hoboken, New Jersey, John Wiley & Sons, 2012, p. 4.

1.1. THE THREE VS IN BIG DATA

The existence of the three Vs in data were first pointed out by a paper that referred to the effects on data management of the surge in e-commerce in the early 2000s[13] but did not mention the phrase "big data". This first element in the conceptualisation of what big data is mostly refers to the characteristics that a dataset must have in order to be labelled as big data.

The first V, volume, refers to the depth and amount of available data.[14] The amount of digitally stored information has undergone a massive increase during the past couple of decades, from a 25% of the world's information being digitally stored in the year 2000 to a 98% in 2013.[15] In addition, this increase in digitally stored information must also be considered in regard to the absolute increase in available information. Hence, in general, there is an unprecedented amount of information available, most of which is digitally stored, meaning it can be processed using the technologies that are the object of this study.

These large quantities of data are created and delivered at a very high speed (velocity) and appear in increasingly different formats (variety). Variety means that the data may come in semi-structured or unstructured formats. A typical example of structured data is an excel sheet that can include information on a series of people, such as their names and contact information.[16] Semi-structured data do not present a defined structure and include much more information than structured formats. Word documents or emails are included within this type of data format. While extracting and organising the information they contain is not as straightforward as it is when presented with tables of data, they do nonetheless offer a certain organisation to them and, consequently, a blueprint for systematising data and obtaining information.[17] For example, e-mail metadata enhances the possibility of classifying the actual information contained the actual message. Unstructured data may appear in many kinds of formats such as text, video or audio. In general, it is much harder to develop relationships

[13] LANEY, D., "3D data management: controlling data volume, velocity and variety", 6th February 2001. Available on 20th September 2018 at: https://blogs.gartner.com/
[14] ARTICLE 29 WORKING PARTY, "Opinion 03/2013 on purpose limitation", 00569/13/EN, WP 203, 2nd April 2013, p. 35.
[15] CUKIER, K. & MAYER-SCHOENBERGER, V., "The rise of big data: how it's changing the way we think about the world", *Foreign Affairs*, vol. 92, No. 93, 2013, p. 29.
[16] TAYLOR, C., "Structured vs. Unstructured data", *Datamation*, 28th March 2018. Available on 13th June 2019 at: https://www.datamation.com/
[17] CONNOLLY, T. & BEGG, C., *Database Systems: A Practical Approach to Design, Implementation, and Management*, Essex, Pearson, 6th ed., 2014, p. 1130.

between pieces of semi-structured or unstructured data than between pieces of structured data. [18]

For example, it is much harder, especially for a machine but also for human beings, to extract the relevant information from a picture than from a table of data. If the objective of an algorithm is to figure out which neighbourhoods different families live in and the only pieces of data available are pictures posted on social media (unstructured data), it will be much more difficult to extract the relevant information from said type of raw data than from an excel document in which the names and addresses of individuals appear (structured data). The similarities between unstructured and semi-structured formats lead some tech experts to only differentiate between structured and unstructured types of data.

1.2. THE NEED TO PROCESS (RAW) BIG DATA

The three Vs make big data highly complex, which leads us to the next part of the explanation: the need to process it. The word 'raw' was used in the previous paragraph to describe big data, which means it must undergo some transformation process to provide useful outputs in the shape of information. One of the key elements of big data is not only that the data must be processed, but that in order to retrieve the relevant information from it, it requires a specific and specially advanced set of methods and technology.[19] Big data has in fact been defined by some as data that cannot be processed using traditional methods and tools.[20] Given its characteristics, special computing power is needed in order to store and process it.

However, placing too much focus on the fact that big data needs special and non-traditional methods and technology would further increase the relativity of the concept for what is now considered non-traditional technology will have probably become normal in just a few years.[21] The important aspect to keep in mind is that big data needs to be processed (by powerful technologies) in order to provide relevant information. This is why the raw quality of the initial datasets that are then processed must be highlighted. Even if a set of facts can be

[18] TAYLOR, C., "Structured vs. Unstructured data", *cit.*, 2018.
[19] DE MAURO, A. *et al.*, "What is big data?...", *cit.*, 2014, pp. 6-7.
[20] GIL GONZÁLEZ, E., *Big data, Privacidad y Protección de Datos*, Madrid, Agencia Española de Protección de Datos, 2016, p. 18.
[21] FRANKS, B., *Taming the big data tidal wave*, *cit*., 2012, p. 4.

extracted from structured datasets, they do not offer knowledge. Knowledge must therefore be obtained by analysing the data with the necessary tools.[22]

1.3. THE FOURTH V: VALUE

The information provided by big data is what brings its conceptualisation full circle. The main reason behind big data becoming such an important phenomenon lies in the value of the information or knowledge it generates,[23] which is why the previous point, referred to data processing and analysis, is so relevant.

The data that is now available is not necessarily better than the data generated by traditional sources;[24] in fact, big data generally contains a lot of irrelevant information before it is processed.[25] However, given the capability of current processing systems, organisations and individuals expect to extract the relevant data from all the noise and thus generate more valuable information even when the initial dataset is not clean.[26] Hence, a trade-off between volume and quality of the raw data is accepted.

Big data processing systems are able to generate quantifiable information on many aspects of the world that had never before been examined using quantitative approaches,[27] such as pictures, videos or other unstructured pieces of information.[28] Even if the same kinds of actions are examined, the data that is now possible to extract from them provides so much more information that it is even possible to qualify big data as a new source of data altogether.[29]

Whether the information generated can actually be considered to be valuable knowledge for the organisation processing the data is something that is fully subjective and will depend on the needs and objectives it has. There are however two general conditions that the information should fulfil. Firstly, the information must be useful for the organisation's

[22] CUSTERS, B., "Data dilemmas in the information society: introduction and overview", in CUSTERS, B., *et al.,* (eds.), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in large Databases*, Berlin, Springer, 2013, p. 8.
[23] DE MAURO, A. *et al.*, "What is big data?…", *cit.*, 2014, pp. 2-3.
[24] FRANKS, B., *Taming the big data tidal wave*, *cit*., 2012, p. 6.
[25] *Ibidem*.
[26] CUKIER, K. & MAYER-SCHOENBERGER, V., "The rise of big data…", *cit.*, 2013, p. 29; LERMAN, J., "Big data and its exclusions", *Stanford Law Review Online,* No. 66, 2013, p. 57.
[27] CUKIER, K. & MAYER-SCHOENBERGER, V., "The rise of big data…", *cit.*, 2013, p. 29
[28] GIL GONZÁLEZ, E., *Big data, Privacidad y Protección de Datos*, *cit.*, 2016, p. 18.
[29] FRANKS, B., *Taming the big data tidal wave*, *cit*., 2012, p. 7.

objectives, which amongst other things means it must provide new knowledge.[30] Additionally, WARD and BARKER point out the need for the outcomes to be trustworthy or, in other words, veracity.[31] In general, only to the extent that the information generated is accurate will it be of any use to the organisation processing big data.

It is important to finally point out the fact that, in many cases, the individual or organisation using the data generated, is not the one processing it. In fact, a whole economic sector known as the data services sector has been created surrounding the new technologies related to big data in order to acquire and process information so as to later sell it on to other companies which then use it to make decisions regarding individuals' chances at getting a loan, insurance or even a job.[32] Thus, the value of the information obtained will largely depend on the needs of the organisation that will use the data.

This difference between the individual or organisation processing the data and the one using the results obtained is reflected in the General Data Protection Regulation (GDPR),[33] which distinguishes between the concepts of "controller", who is "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data",[34] and "processor", who is "a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller".[35] This terminology will be used throughout the thesis to identify those responsible for data processing systems.

2. DATA PROCESSING TOOLS AND TECHNOLOGIES

2.1. MACHINE LEARNING AND DATA MINING

In order to extract the relevant knowledge from big data it is necessary to use a specific set of tools amongst which data mining and machine learning are included. Data mining is part of the knowledge discovery in databases process. Knowledge discovery in databases can be

[30] CUSTERS, B., "Data dilemmas in the information society: introduction and overview", *cit.*, 2013, p. 9.
[31] WARD, J. & BARKER, A., "Undefined by data...", *cit.*, 2013, p. 1.
[32] US EXECUTIVE OFFICE OF THE PRESIDENT, "Big data: seizing opportunities, preserving values", 2014, pp. 43-44.
[33] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
[34] Article 4.7 GDPR.
[35] Article 4.8 GDPR.

defined as "the overall process of discovering useful knowledge from data"[36]. Within this process, data mining constitutes the step in which relevant relationships are extracted from the data. Data mining can be thus described as the part of the process of analysis in which algorithms are used to discover patterns in the data that would probably not be detected by human analysts.[37]

According to HAND, who already defined data mining in 1998, it is "the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners".[38] Data mining is thus used to extract rules from the available data by obtaining implicit, previously unknown and potentially useful information.[39]

Machine learning is a subfield of computer science, which includes a set of processes or methods that can find existing correlations in datasets, and use the discovered patterns in order to make predictions.[40] Machine learning systems are able to constantly learn and improve over time through the use of techniques such as neural networks, which connect ideas in similar ways to human brains.[41] These systems can develop as far as making 'intelligent' decisions similar to those that a human being in the same position would have made, thus being generally considered a branch of Artificial Intelligence.[42]

Machine learning and data mining tend to work together,[43] seeing as the algorithms used to extract relevant relationships during the data mining phase are generally machine learning algorithms.[44] It is quite difficult to establish a difference between data mining and machine

[36] FAYYAD, U. M., PIATESKY-SHAPIRO, G. & SMYTH, P., "From data mining to knowledge discovery in databases", *AI Magazine*, vol. 17, No. 3, 1996, p. 39.
[37] HILDEBRANDT, M. & KOOPS, B. J., "The challenges of ambient law and legal protection in the profiling era", *The Modern Law Review*, vol. 73, No. 3, 2010, pp. 431-432.
[38] HAND, D. J., "Data mining: statistics and more?", *The American Statistician*, vol. 52, No, 2, 1998, p. 112.
[39] SAHU, H., SHRMA S., & GONDHALAKAR, S., "A brief overview on data mining survey" *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, vol. 1, No. 3, 2013, p. 114.
[40] MURPHY, K. P., *Machine Learning: A Probabilistic Perspective*, Cambridge (Massachusetts), The MIT Press, 2012, p. 1.
[41] TEGMARK, M., *Life 3.0. Being Human in the Age of Artificial Intelligence*, London, Penguin Books, 2017, pp. 97-107.
[42] SURDEN, H., "Machine learning and law", *Washington Law Review,* vol. 89, 2014, pp. 89-90.
[43] WITTEN, I. H. *et al.*, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Cambridge (Massachusetts), Morgan Kaufman, 2017, p. 28.
[44] MURPHY, K. P., *Machine Learning…. cit.*, 2012, p. 1: "Most closely related to data mining is without doubt machine learning. There is a big overlap between the two communities, and over time the difference became less relevant and boundaries are beginning to blur. Traditionally, machine learning is about learning to perform a task, whereas data mining is more about "finding knowledge from the data". Both are tightly connected; on the one hand, in general, useful knowledge extracted from given examples of a task will allow for performing the task better, whereas on the other hand, during the learning process of a task, knowledge about the task will have to be accumulated in one form or another, from the examples, and be stored in the system. Given its task-

learning since both tools are used in order to extract relevant relationships and make predictions.[45] One of differences is that, while data mining mostly focuses on data collection and establishing relationships, machine learning is used in order to predict future outcomes and make decisions based on that information.[46] Another key difference, one that becomes essential in dealing with the discriminatory outcomes of automated decision-making systems, is that, in data mining the models produced are more interpretable but less accurate than machine learning models.[47]

One of the most relevant aspects of machine learning is that the algorithm will keep learning on its own as it receives more and more information even when the model has already been deployed.[48] An example in which machine learning is used is music streaming platforms such as Spotify in which the programme will be able to detect the user's music tastes and thus offer her different playlists according to said music taste.[49] Applications such as Spotify or Netflix, which offer recommendations, use a specific subset of machine learning tools known as deep learning[50] which, due to their effectiveness and efficiency are increasingly used in different areas such as medicine[51] or manufacturing.[52]

2.2. SUPERVISED AND UNSUPERVISED LEARNING

The biggest difference between general machine learning systems and the subset of deep learning tools is the degree to which the system needs instructions. Programmes do end up developing knowledge on their own in both cases. However, in more traditional machine learning systems, this autonomous development is heavily directed by coding, especially in

---

oriented nature, historically one can see the ML community having a strong focus on supervised tasks, whereas data mining is more concerned with unsupervised tasks."

[45] OQUENDO, M. A. *et al.*, "Machine learning and data mining: strategies for hypothesis generation", *Molecular Psychiatry*, vol. 17, No. 10, 2012, p. 957.

[46] MURPHY, K. P., *Machine Learning…. cit.*, 2012, p. 1.

[47] *Idem*, p. 16.

[48] TEGMARK, M., *Life 3.0. Being human in the age of Artificial Intelligence*, *cit.*, 2017, p. 97.

[49] GROSSFELD, B., "A simple way to understand machine learning vs deep learning", Zendesk, 18th July 2017. Available on 31st January 2019 at: https://www.zendesk.com/

[50] LIU, J. & WU, C., "Deep learning based recommendation: a survey" in KIM K. & JOUKOV N. (eds), *Information Science and Applications 2017. ICISA 2017*, Lecture Notes in Electrical Engineering, vol. 424, Singapore, Springer, p. 452; COGLIANESE, C. & LEHR, D., "Regulating by robot: administrative decision making in the machine-learning era", *The Georgetown Law Journal*, vol. 105, No. 5, 2017, p. 1160.

[51] LEE, J. G. *et al.*, "Deep learning in medical imaging: general overview", *Korean Journal of Radiology*, vol. 18, No. 4, 2017, pp. 570-584.

[52] WANG, J. *et al.*, "Deep learning for smart manufacturing: methods and applications", *Journal of Manufacturing Systems*, vol. 48, 2018, pp. 144-156.

the training stages.[53] Conversely, in the case of deep learning, the computer is able to draw conclusions, develop knowledge and tune itself with very little instructions.[54] In these cases, the machine is coded so that it is mostly autonomous and can learn from its surroundings, reaching conclusions in a very similar manner to the way a human brain works.[55]

These new technologies can be used in order to confirm suspected correlations between variables. In fact, algorithms are constantly put to use in order to verify relationships between data.[56] This kind of analysis is known as top-down or supervised learning and is much closer to traditional statistical analysis[57] as it requires feeding the computer with a selected sample of data in order for it to extract the relevant relationships. This initial set of data is known as training data and it needs to be collected and prepared before it is processed.

However, a different way in which machine learning can be put to use is through bottom-up analysis, also known as unsupervised learning. These techniques greatly differ from traditional statistical analysis since, instead of developing a hypothesis, which is then tested over and over again with the available data; data is fed into a computer programme, which then extracts the relevant hypotheses.[58]

The inversion of the traditional process is made possible due to the vast amount of data that is now available and the development of the necessary technologies to process it. The enormous volume of data makes it very hard for human beings to be able to detect the possible relationships in it, thus rendering the use of automated systems necessary. In addition, the availability of such large quantities of data in theory ensures that the data processing computer programmes will reach more accurate results than when processing smaller datasets.[59]

Through the following example, we aim to illustrate the difference between supervised and unsupervised learning. Imagine a supermarket chain believed that there was a group of employees who were stealing cleaning products and it asked the firm's in-house IT

[53] BOSTRON, N., *Superintelligence. Paths, Dangers, Strategies*, Oxford University Press, Oxford, 2014, pp. 179-180.
[54] GROSSFELD, B., "A simple way to understand machine learning vs deep learning", *cit.*, 2017.
[55] *Ibidem*.
[56] HILDEBRANDT, M. & KOOPS, B. J., "The challenges of ambient law and legal protection in the profiling era", *cit*., 2010, p. 432.
[57] *Ibidem*.
[58] CUSTERS, B., "Data dilemmas in the information society: introduction and overview", *cit.*, 2013, p. 7.
[59] O'NEIL, C., *Weapons of Math Destruction…*, *cit.*, 2017, p. 6.

department to develop an algorithm to verify this suspicion and identify the responsible parties. The individuals developing the system would select a representative sample of employees and inform the algorithm of what behaviour is considered stealing and what behaviour is not. In other words, the data provided would be labelled. Then, the company would follow the daily work-routine of the selected sample of employees so as to determine which staff members were stealing and which were not. This information would be improved by other data such as employee schedules, positions in the company's hierarchy, their behaviour in the moments before and after stealing and days and times in which the products were stolen, amongst others.

All the information gathered would enable the firm to estimate the cost of stolen cleaning products for the selected sample and extrapolate it to the entire firm. More importantly, by feeding this information into a machine learning algorithm, the supermarket chain would be able to build a model to predict which employees would steal cleaning products and flag them so that enhanced supervision was placed on suspicious workers.

For the bottom-up or unsupervised approach, all of the data regarding employees, their behaviours and schedules would be introduced into the machine learning algorithm in order for it to figure out any relevant relationships, not just with regard to the stealing problem but to any issue, such as job performance. In this case the historical data fed to the algorithm would not be labelled and the relationships between data and results would be completely unknown.

Hence, the key element in supervised learning is that the information that is being searched for is already known and labelled; this enables the organisation to feed the algorithm with already selected and partly pre-processed data.[60] Conversely, unsupervised learning is used in order to extract completely unknown and unsuspected relationships in the data, leading to a reduction in the human control of the process.[61] Nonetheless, although the information that is being searched for is already known in supervised learning, which undoubtedly facilitates the implementation of tools aimed towards controlling these systems, algorithms and models using supervised learning are still highly opaque.

---

[60] HILDEBRANDT, M. & KOOPS, B. J., "The challenges of ambient law and legal protection in the profiling era", *cit.*, 2010, p. 432.

[61] HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Berlin, Springer, 2009, p. xi.

Supervised and unsupervised learning can be complementary seeing as once the subjects have been clustered according to the relationship between their input variables; these groups can be used to develop supervised analysis.[62] Supervised learning algorithms are used in applications such as predictive policing, credit scoring, and predicting employee performance and thus currently have greater implications from a legal perspective. [63]

2.3. ALGORITHMS AND MODELS

Throughout the dissertation, automated decision-making systems will be referred to in many instances as machine learning models and algorithms. As O'NEIL puts it, models are "nothing more than an abstract representation of some process". [64] An algorithm is the list of instructions for analysing the data in order to provide the user with a certain answer or result given the available information or, in other words, a software code that processes a limited amount of instructions.[65] Machine learning models and algorithms are "pattern recognition tools"[66] which are trained with and use data mining and machine learning tools in order to extract or predict information regarding a certain phenomenon.

In supervised learning, algorithms (lists of instructions) are trained by being "punished" or "rewarded" depending on how accurate their predictions for the training data are,[67] thus refining their interpretative and predictive capabilities. The algorithm will try different predictive rules (lists of instructions), finally choosing one that, according to the information the algorithm is provided, yields the most accurate predictive results for the historical data.[68]

For example, if the algorithm is being trained to distinguish between numbers when they are hand-written, it will be fed correctly labelled images of numbers so that it knows when it is making errors.[69] Once it has been trained on historical data, it generates a learning tool that will be the final product used by organisations in order to feed it new data that it will process

[62] LEHR, D. & OHM, P., "Playing with the data: what legal scholars should learn about machine learning", *UC Davis Law Review,* vol. 51. No. 2, 2017, p. 676.
[63] LEHR, D. & OHM, P., "Playing with the data…" *cit.*, 2017, p. 676.
[64] O'NEIL, C., *Weapons of Math Destruction…*, *cit.*, 2017, p. 18.
[65] MONASTERIO ASTOBIZA, A., "Ética algorítmica: Implicaciones éticas de una sociedad cada vez más gobernada por algoritmos", *Dilemata,* No. 24, 2017, p. 185.
[66] SUTHAHARAN, S., *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, New York, Springer, 2015, p. 123.
[67] COGLIANESE, C. & LEHR, D., "Regulating by robot…", *cit.*, 2017, p. 1158.
[68] LEHR, D. & OHM, P., "Playing with the data…" *cit.*, 2017, pp. 671-672.
[69] COGLIANESE, C. & LEHR, D., "Regulating by robot…", *cit.*, 2017, p. 1158.

and produce results on.[70] Regarding the uses of these technologies that will be analysed through the dissertation, what generally results from this process is a model, an automated decision-making system that represents real life decision-making processes. The model will be the final tool that is deployed in order to make real decisions.

Models can be created without using algorithms and algorithms can be completely abstract and not be used to model anything. However, given the fact that the dissertation focuses on automated decision-making, that is, using these technologies to model human decision-making, the systems analysed contain algorithms as sets of instructions which provide information the way in which the decision-making model must behave.

If a model is designed to predict who will default a mortgage payment, during the training process, the technologies used to analyse the training data might have discovered that unemployed people with more than one child under the age of fifteen are more likely to default. This is a rule that the deployed model will use when it is fed data on new subjects to determine whether they will default. Therefore, during training, learning techniques are used to find the relevant relationships between the data that will then serve as rules for when the model is processing real-world data.[71]

The objective of this section has been to draw a very general picture of the amount of information that these systems can process and generate and the degree of autonomy and complexity they can achieve so as to illustrate how the difficulty of understanding and controlling these systems (which clearly go further beyond simply cross-referencing information in two databases) generates the need for new regulatory instruments that specifically address algorithmic decision-making.

3. THE APPLICATION OF AUTOMATED SYSTEMS

The technologies that have been described in the previous section have enabled both public and private organisations to take automated decision-making to the next level. The amount of data that can now be processed and the capacity that machine learning tools have to extract relevant information has allowed for the development of much more complex decision-

[70] *Ibidem*.
[71] Throughout the dissertation, the technologies analysed with regard to existing legal frameworks, and for which a new regulatory framework shall be proposed, are mainly referred to as algorithms and automated systems and, in some cases, data processing technologies, models and software programmes.

making systems. Hence, on top of the simple decision-making systems that have already been put in place for many years, a large number of possibilities have opened up, enabling the public and private sector to maximise process efficiency.[72]

As the following pages will convey, algorithms and other technological related applications have become an essential part of many aspects of modern Western societies.[73] The reason why algorithms have become so prevalent and are being used for many different purposes is the rate at which technology, and particularly, computational capacity, is evolving.[74] Automated processes are being constantly and very rapidly developed and improved and are easily introduced in environments, such as smart cities, in which the use of these systems is becoming increasingly common.[75] Thus, to a certain extent, it can be argued that algorithms produce a series of benefits that outweigh the risks and costs generated by automatisation. Nonetheless, in some cases, it does seem like the risks and costs of automated systems are not being sufficiently taken into consideration by the individuals and organisations employing them.

The rapid integration of these technologies in both the private and public sector brings about a series of concerns regarding the paradigm shift many areas of society will undergo. Particularly, with regard to the public sector, the use of machine learning models by public institutions unavoidably forces us to rethink the principles of democratic governance and administrative law,[76] especially if algorithms are being used in public decision and rule-making.[77]

As the examples explained in the following pages convey, algorithms increasingly impact human lives,[78] generating a greater sense of urgency regarding the need for regulation. In fact, most of the reports published by public institutions have focused on the current and future social impact of algorithms.[79] The social implications that automated systems may

[72] PADOFF, R., "Why deep learning is suddenly changing your life", *cit.*, 2019.
[73] MONASTERIO ASTOBIZA, A., "Ética algorítmica…", *cit.*, 2017, p. 188.
[74] CATH, C. *et al.*, "Artificial intelligence and the 'good society': the US, EU and UK approach", *Science and Engineering Ethics*, vol. 24, No. 2, 2018, p. 506.
[75] RANCHORDÁS, S., "Nudging citizens through technology in smart cities", *International Review of Law, Computers & Technology,* vol. 33, 2019, pp. 1-23.
[76] COGLIANESE, C. & LEHR, D., "Regulating by robot…", *cit.*, 2017, p. 1152-1153.
[77] *Idem*, p. 1180.
[78] MONASTERIO ASTOBIZA, A., "Ética algorítmica…", *cit.*, 2017, p. 188.
[79] EUROPEAN PARLIAMENT, "European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics" 2015/2103(INL), 2017; US EXECUTIVE OFFICE OF THE

have, range from the consequences that a substitution of the workforce by machines may have,[80] to the new challenges that are already arising regarding the protection of fundamental rights and freedoms. While all the risks generated by algorithms are intimately related, this dissertation mostly focuses on the harms they cause to fundamental rights and freedoms and, especially, to the rights to equality and non-discrimination.

The following pages offer a general overview of the ways in which algorithms are used in both the public and private sectors in order to convey the degree of penetration of complex automated decision-making systems in different areas and also to highlight some of the uses that can generate a significant impact on individuals' rights in a direct or indirect manner.

3.1. THE USE OF ALGORITHMS BY THE PRIVATE SECTOR

The use of algorithms in the private sector is widely extended, thus making it an impossible task to cover all of their possible applications. Some of the most commonly known applications of algorithms are spam filters,[81] search engines[82] and autonomous cars.[83] The following examples cover some of the uses of algorithms in decisions that most directly affect individuals' lives, some of which will later be discussed in more detail when addressing the discriminatory effects of automated decision-making.[84]

PRESIDENT, "Artificial intelligence, automation and the economy", 2016; UK GOVERNMENT OFFICE FOR SCIENCE, "Artificial intelligence: an overview for policymakers", 2016.

[80] The substitution of human workers, will mostly affect low and middle income workers The changes caused by automation entail the need for a highly specialised workforce trained in highly technical skills, meaning governments will have to design employment and education policy accordingly in order to avoid a massive increase in economic equality resulting from a surplus in unspecialised workers whose skills are no longer required in the digital and automated economy. US EXECUTIVE OFFICE OF THE PRESIDENT, "Artificial intelligence, automation and the economy", *cit.*, 2016, pp. 13-21 and 26; COGLIANESE, C. & LEHR, D., "Regulating by robot…", *cit.*, 2017, p. 1150; CATH, C. *et al.*, "Artificial intelligence and the 'good society'…", *cit.*, 2018, p. 510.

[81] GUZELLA, T. S. & CAMINHAS, W. M., "A review of machine learning approaches to spam filtering", *Expert Systems with Applications*, vol. 36, No. 7, pp. 10206-10222.

[82] DAVIES, D., "How search engine algorithms work: everything you need to know", *Search Engine Journal*, 10th May 2018. Available on 3rd April 2019 at: https://www.searchenginejournal.com/

[83] Autonomous cars have become a very prominent topic in discussions regarding automation due to the ethical dilemmas that arise regarding the instructions the car should follow in cases in which it has to choose between putting either the passenger or a pedestrian's life at risk. See, for example, RENDA, A., "Ethics, algorithms and self-driving cars – a CSI of the 'trolley problem'", *CEPS Policy Insights,* No. 2018/02, 2018, pp. 1-15.

[84] The role of search engines and other website algorithms will also be discussed in chapter three in order to convey the role they have in promoting discriminatory attitudes and perpetuating negative stereotypes. However, given that this chapter mainly focuses on the pervasiveness of algorithms in society the examples analysed in more detailed are those of algorithmic applications which generate decisions that affect individuals more directly.

3.1.1. Scoring individuals

In the United States, credit-reporting agencies have, for a few decades now, been collecting information on consumers. This data is then used by all kinds of firms in order to determine their eligibility for, amongst other things, jobs, loans or insurance.[85] For example, information on whether a loan applicant has paid her bills in time over the past few years, has any other form of debt, has a savings account and any other relevant details is collected in order to draw a credit profile for the individual.[86] The resulting information is then compared to reports on other individuals with similar profiles and the creditworthiness score of the applicant is determined.[87]

The amount of information which is currently available makes it possible for firms to score individuals based not solely on their finances but also on many other elements which in theory allow for the creation of more accurate scores that can be adjusted depending on the objective at hand,[88] thereby expanding the role of automated decision-making in all sectors.

3.1.1.1. The banking sector and the expansion of credit scores

Algorithms have been used by credit card companies[89] and the banking sector in the US for quite some time. The main purpose of these algorithms is to establish individuals' credit capacity and make decisions on their eligibility for loans. These decisions have been based on creditworthiness scores for around the past six decades.[90] Before credit scores entered the picture, it was bank employees and, later on, experts, who decided what someone's trustworthiness ought to be and whether the loan they had requested should be granted or denied.[91] Eventually, specialised companies began creating models to detect the probability

[85] US EXECUTIVE OFFICE OF THE PRESIDENT, "Big data…", *cit.*, 2014, p. 44.
[86] KROLL, J. *et al.*, "Accountable algorithms", *University of Pennsylvania Law Review*, vol. 165, No. 3, 2017, p. 658.
[87] FURLOW, B., "IBM Watson collaboration aims to improve oncology decision support tools", *The Journal of Oncology*, 16th March 2016. Available on 3rd April 2019 at: https://www.cancernetwork.com/; US EXECUTIVE OFFICE OF THE PRESIDENT, "Big data…", *cit.*, 2014, p. 45.
[88] PASQUALE, F., *The Black Box Society: The Secret Algorithms that Control Money and Information*, Cambridge (Massachusetts), Harvard University Press, 2015, p. 25.
[89] US District Court for the Northern District of Georgia, Atlanta division, "Complaint for permanent injunction and other equitable relief at 35 FTC v. Compucredit Corp", No. 1:08-CV-1976-BBM, 2008.
[90] ABDOU, H. A., & POINTON, J. "Credit scoring, statistical techniques and evaluation criteria: a review of the literature", *Intelligent Systems in Accounting, Finance & Management,* vol. 18, No. 2-3, 2011, pp. 59-88.
[91] CITRON, D. K. & PASQUALE, F, "The scored society: due process for automated predictions", *Washington Law Review Online*, vol. 89, 2014, p. 8.

that a loan applicant would default when repaying the bank, thereby producing credit scores.[92]

Through the development of new big data technologies, the sources of information employed in building individuals' credit scores have increased exponentially. Moreover, new versions of these scores are now adapted to the specific needs of a wide range of fields,[93] from human resources to car insurance.[94]

It is important to highlight that the United States is not the only country in which these reputational systems for determining creditworthiness are used. Credit scoring is, for instance, also used in the United Kingdom[95] and Canada[96] as well as in several Asian countries such as Malaysia, Singapore and Hong Kong[97] and is rapidly expanding throughout other parts of the world.

In European countries other than the UK, loans have traditionally been granted relying on a reduced number of elements, such as salary, job security or family financial situation, which determine the probability that a loan applicant would default in his or her payments.[98] More complex scoring systems have, however, gradually entered the banking sector and are now also widely used[99] although mostly limited to loan granting and not with regard to other banking services.

3.1.1.2. Healthcare

The use of big data and predictive models in healthcare undoubtedly has a great deal of positive effects seeing as, by using these new technological developments, it is now possible

[92] *Idem*, p. 9.
[93] In fact, even a dating website takes credit scores as the main premise when matching customers who hire their services. See CREDIT SCORE DATING. Available on 27th March 2019 at: www.creditscoredating.com
[94] CONSUMER REPORTS, "The secret score behind your auto insurance", 10th August 2006.
[95] BALL, K., "Blacklists and black holes: credit scoring in Europe", in WEBSTER, W., & BALL, K., (eds.), *Surveillance and Democracy in Europe: Courting Controversy*, Oxon, Routledge, 2019, p. 69.
[96] PAYNE, A., "Credit score systems across the world", *Graydon*, 9th February 2015. Available on 25th February 2019 at: https://www.graydon.co.uk/
[97] *Ibidem*.
[98] KAYNE, C., "Do credit scores matter outside the US?", *CNBC*, 9th February 2011. Available on 25th February 2019 at: https://www.cnbc.com/
[99] BALL, K., "Blacklists and black holes…", *cit.*, 2019, p. 71; BBVA, "How credit scoring can influence the granting of a loan". Available on 25th February 2019 at: https://www.bbva.es/

to know a disease will appear before it does and design personalised plans in order to prevent it or, at least, reduce its impact on the patient's health.[100]

Private insurers manage an important part of healthcare. The extent to which these companies intervene in healthcare depends on the size of the welfare state in each country. However, even in countries such as Spain or the UK in which public healthcare is still predominant, there is a clear tendency to increasingly privatise medical services.[101]

The use of predictive medicine by healthcare insurance companies is gradually expanding due to the technological developments which allow, amongst other things, to predict a person's chances of contracting certain illnesses,[102] and adapt insurance primes accordingly. The use of algorithms in this sector is also largely the consequence of the amount of health-related information that is now available following the commercialisation of quantified-self devices and apps.[103] These devices allow medical insurers to evaluate individuals' general lifestyle, habits and health in order to determine whether they are at risk or have a certain propensity for needing or demanding a significant number of medical services.

3.1.1.3. Human resources

Companies increasingly rely on algorithms when recruiting for job openings. In fact, in a 2017 survey carried out by a recruitment software company, around 55% of the firms asked said that by 2022 at least part of their recruiting processes would be automated.[104] Using the new technologies available makes hiring more effective and efficient given that algorithms can be programmed to select candidates whose CV best fits the job description.[105] In

[100] BURT, A. & VOLCHENBOUM, S., "How healthcare changes when algorithms start making diagnoses", *Harvard Business Review*, 8th March 2018. Available on 25th March 2019 at: https://hbr.org/

[101] See, in general, SÁNCHEZ-MARTÍNEZ, F. I., ABELLÁN-PERPIÑÁN, J. M. & OLIVA-MORENO, J., "Privatization in healthcare management: an adverse effect of the economic crisis and a symptom of bad governance. SESPAS report 2014", *Gaceta Sanitaria*, vol. 28, No. 1, 2014, pp. 75-80.

[102] CHEN, M. *et al.*, "Disease prediction by machine learning over big data from healthcare communities", *IEEE Access*, vol. 5, 2017, pp. 8869-8879.

[103] MALGIERI, G. & COMANDÉ, G., "Sensitive-by-distance: quasi-health data in the algorithmic era", *Information & Communications Technology Law*, vol. 26, No. 3, 2017a, p. 231.

[104] CAREERBUILDER, "More than half of HR managers say artificial intelligence will become a regular part of HR in next 5 years", 18th May 2017. Available on 12th February 2019 at: https://www.prnewswire.com/

[105] FALIAGKA, E. *et al.*, "On-line consistent ranking on e-recruitment: seeking the truth behind a well-formed CV", *Artificial Intelligence Review*, No. 42, 2014, p. 516; FALIAGKA, E., RAMANTAS, K. & TSAKALIDIS, A., "Application of machine learning algorithms to an online recruitment system", paper presented at the *7th International Conference on Internet and Web Applications and Services*, 2017. Available on 26th March 2019 at: http://citeseerx.ist.psu.edu/

addition, firms can now use consumer profiles in order to send job advertisements, targeting only those people who are better suited for the position offered.[106]

Algorithms are also used in order to evaluate and predict employees' future performance,[107] which can be very helpful when firms make restructuring decisions, when employee evaluations are carried out, and for addressing organisational issues.[108]

3.1.2. Consumer profiling and advertising

Algorithms are used in order to create consumer profiles, which can be used in order to market products and services in a more efficient way.[109] As the amount of information available to data brokers increases, so does the detail with which consumer profiles can be drawn. Data brokers use online information from purchases, social network profiles, general network interactions, customer support and combine it with other information from public records in order to draw very accurate descriptions of consumers, fitting them into specific categories.[110]

3.2. THE USE OF ALGORITHMS BY THE PUBLIC SECTOR

Although the public sector has been using algorithms for a very long time,[111] the development of machine learning tools, which improve the effectiveness and efficiency of many different types of tasks, has led to their increased use by public administrations and organisations in general.[112] From automated decision-making systems that identify which restaurants in a city should be inspected,[113] or automatising aid systems,[114] to algorithms that

[106] BOGEN, M. & RIEKE, A., "Help wanted: an examination of hiring algorithms, equity and bias", *Upturn*, p. 17.
[107] KIRIMI, J. M. & MOTURI, C. A., "Application of data mining classification in employee performance prediction", *International Journal of Computer Applications*, vol. 146, No. 7, 2016; PECK, D., "They're watching you at work", *The Atlantic*, December 2013. Available on 20th February 2019 at: https://www.theatlantic.com/
[108] EY, "The new age: artificial intelligence for human resource opportunities and functions", 2019. Available on 3rd April 2019 at: https://www.ey.com/
[109] BAR-GILL, O., "Algorithmic price discrimination when demand is a function of both preferences and (mis)perceptions", *Chicago Law Review*, vol. 86, No. 2, 2019, p. 219.
[110] VALLETTI, T., & WU, J., "Consumer profiling with data requirements", *Production and Operations Management*, Vol. 29, No. 2, 2020, pp. 309-329; US EXECUTIVE OFFICE OF THE PRESIDENT, "Big data…", *cit*, 2014, p. 44.
[111] BING, J., "Code, access and control", *cit.*, 2005, pp. 203-204.
[112] COGLIANESE, C. & LEHR, D., "Regulating by robot…", *cit.*, 2017, p. 1161.
[113] *Ibidem*.
[114] EUBANKS, V., *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, New York, St Martin's Press, 2017.

nudge[115] citizens into making certain decisions,[116] automation has become an essential part of some of the activities carried out by governments and will be increasingly prevalent in the future.[117]

The examples provided in the following pages are classified according to the division of public administration's actions proposed by SANTI ROMANO, who inspired his theory in the differentiation between public service provision and limitative activities that had been previously put forward by the German scholarship.[118] Public limitative activities encompass all types of regulation, from business permits to police activity.[119] This type of public activity is thus also labelled regulatory, coercive or police activity. Examples of public aid allocation through algorithms are included under the use of automation in public service management and provision.

3.2.1. The use of algorithms in public service management and provision

Algorithms can be used to improve the lack of efficiency and effectiveness that have traditionally characterised some of the actions carried out by public administrations. The concerns that arise from this lack of efficiency and effectiveness are particularly problematic in cases in which social services should urgently intervene. With this objective in mind, several algorithms have been developed, for example, so that social services can rapidly address children welfare cases.[120]

In the health sector, medical data, including a person's genetic information and exercise or dietary habits, may be collected and used in order to, amongst other objectives, determine an individual's chance of developing a certain disease or illness.[121] For example, a public

[115] See, in general, SUNSTEIN, C. R. Y THALER, R. H., *Nudge: Improving Decisions about Health, Wealth and Happiness*, New Haven, Yale University Press, 2008.
[116] RANCHORDÁS, S., "Nudging citizens through technology in smart cities", *cit.*, 2019, p. 18.
[117] COBBE, J., "Administrative law and the machines of government: judicial review of automated public-sector decision-making", *Legal Studies*, vol. 39, No. 4, 2019, p. 654.
[118] COSCULLUELA MONTANER, L., *Manual de Derecho Administrativo*, Cizur Menor (Aranzadi), 27th ed., 2017, p. 604; MUÑOZ MACHADO, S., *Tratado de Derecho Administrativo y de Derecho Público General. Tomo XIV. La Actividad Regulatoria de la Administración*, Madrid, Boletín Oficial del Estado, 2015, p. 14.
[119] GARRIDO FALLA, F., "El concepto de servicio público en el derecho español", *Revista de Administración Pública*, No. 135, 1994, p. 20.
[120] ECKERD CONNECTS, "Eckerd rapid safety feedback". Available on 4th April 2019 at: https://eckerd.org/; KITZMILLER, E. M., "IDS case study: Allegheny County. Allegheny County's data warehouse: leveraging data to enhance human service programs and policies", *Actionable Intelligence for Social Policy*, May 2014. Available on 4th April 2019 at: https://www.aisp.upenn.edu/
[121] US EXECUTIVE OFFICE OF THE PRESIDENT, "Big data…", *cit.*, 2014, p. 23.

hospital in Spain has developed an algorithm which improves the early detection of sepsis cases.[122]

Furthermore, and although the focus of this research is the role of algorithms in Western societies, it is also relevant to point out how the use of these technologies in developing countries can prevent the national emergency situations that are sometimes caused by disease outbreaks. A clear example is the development of an algorithm used in Cambodia to promptly detect major dengue outbreaks.[123]

Moreover, the use of algorithms has proven useful in the recent Covid19 outbreak. Automated systems have been and are, at the time of writing, being used in order to track the virus, detect especially vulnerable patients and for other purposes aimed towards improving management of the virus.[124]

Algorithms and models are also used in order to improve public education, especially in the USA. For example, O'NEIL[125] analyses the case of teacher evaluations in Washington, D.C. The evaluation resulted from the combination school administrators' opinion on each teacher with a score generated by algorithm. The automated score outweighed opinions in order to reduce human bias. This particular case became especially prominent given that no procedural rights were granted to teachers fired as a result of their algorithmic score. Additionally, it is also suspected that the system was not able to detect inflated test scores resulting from teachers correcting their students' answers in standardised tests.[126]

In addition, the US Department of Education has been working on improving teaching and learning through the use of data mining and learning analysis.[127]

[122] REDACCIÓN MÉDICA, "'Big data' e IA mejoran un 40% la detección precoz de la sepsis grave", 10th March 2019. Available on 3rd April 2019 at: https://www.redaccionmedica.com/
[123] LEDIEN, J. *et al.*, "An algorithm applied to national surveillance data for the early detection of major dengue outbreaks in Cambodia", *PLOS One*, vol. 14, No. 2, 2019, pp. 1-11.
[124] ALIMADADI, A. *et al.*, "Artificial intelligence and machine learning to fight COVID-19", *Physiological Genomics*, vol. 52, 2020, pp. 200-202.
[125] O'NEIL, C., *Weapons of Math Destruction…*, *cit.*, 2017, pp. 3-11.
[126] *Ibidem*.
[127] US DEPARTMENT OF EDUCATION, "Enhancing teaching and learning through educational data mining and learning analysis", October 2012.

3.2.1.1. The use of algorithms in public aid and welfare programmes

Many countries have chosen to automate their welfare programmes in order to allocate resources more efficiently and save money by preventing and detecting cases of fraud. Automating public aid and welfare programmes, raises problems with regard to the prioritisation of economic efficiency over redistribution objectives in this particular area of public activity. However, since this chapter is exclusively focused on conveying the generalised and increasing use of algorithms, said issues will not be addressed here but later on in the dissertation. An especially widespread use of algorithms for these purposes has taken place in child protection services seeing as several countries, such as Denmark, New Zealand, the United Kingdom and the United States, are using algorithms to detect and classify children at risk.[128]

Algorithms are also being used to determine public aid beneficiaries. The Municipal Housing and Land Company in the city of Madrid began using an algorithm to allocate public housing in 2019.[129] All housing petitioners are divided into several demand groups, which, for example, include single parent families; people with disabilities; individuals under 35, and people in an extreme situation of social exclusion amongst others.[130]

Most of the groups into which petitioners are classified in can be labelled as vulnerable or at risk of not being able to afford housing. There is an additional group labelled "general demand", which encompasses petitioners that do not fall under any of the vulnerable groups. This initial classification is carried out by public servants and not by the algorithm. Within each group, subgroups can also be elaborated. These subgroups consider the number of minors, people with a disability and the income level of each household. Once the groups and subgroups have been defined, the algorithm randomly allocates available housing spaces within each group or, if it is the case, within each subgroup.

In Spain, the process of determining eligibility for a discount in the energy bill has also been automated. This particular type of public aid programme is known as the "energy social bond". The use of this algorithm has become especially controversial since there have been

[128] ALSTON, P., "Digital welfare states and human rights", UN Special Rapporteur on extreme poverty and human rights, report A/74/493, A/74/493, 11th October 2019, pp. 10-11.
[129] EMPRESA MUNICIPAL DE LA VIVIENDA Y EL SUELO, "Procedimiento de adjudicación", *Ayuntamiento de Madrid*. Available on 27th November 2019 at: https://www.emvs.es/
[130] Regulation 20th December 2018, for the adjudication of housing managed by the municipal housing and land company of Madrid (article 13).

several cases of individuals who complied with the requirements set by the regulation but were still denied the corresponding discount.[131] Although a civil society organisation has requested access to the algorithm, said access has been denied both by the Spanish government and the Transparency Council, and the case is currently pending before the Spanish Central Administrative Court.[132]

3.2.2. The use of algorithms in public administration's regulatory and coercive activity: law enforcement

3.2.2.1. Police departments and the criminal justice system

The use of machine learning algorithms has become especially widespread amongst police departments and the criminal justice system given the amount of data that is available to them.[133] The fact that this form of automated decision-making affects the fundamental rights of individuals, means that it must be employed with special caution.

OSWALD and GRACE[134] establish three categories of law enforcement algorithms. The first category comprises systems used to detect hotspots in which more criminal activity is likely to be carried out.[135] Detecting areas in which criminal offences are likely to occur provides police departments with the possibility of allocating their resources more efficiently and with a wider timeframe in which to plan and organise police activity.[136]

Algorithms are also used in law enforcement in order to predict specific threats and analyse the data available in on-going criminal investigations.[137] The use of automated systems for these purposes provides an insight into specific situations that may be overlooked by law enforcement officials, such as connections between victims or criminals.[138]

---

[131] BELMONTE, E., "La aplicación del bono social del Gobierno niega la ayuda a personas que tienen derecho a ella", *CIVIO*, 16th May 2019. Available on 6th December 2019 at: https://civio.es/

[132] CIVIO, "Que se nos regule mediante código fuente o algoritmos secretos es algo que jamás debe permitirse en un Estado social, democrático y de Derecho", *CIVIO*, 2nd July 2019. Available on 6th December 2019 at: https://civio.es/

[133] BELLOVIN, S. M. *et al.*, "When enough is enough: location tracking, mosaic theory, and machine learning", *NYU Journal of Law & Liberty*, vol. 8, 2014, p. 612.

[134] OSWALD, M. & GRACE, J., "Intelligence, policing and the use of algorithmic analysis: a freedom of information-based study", *Journal of Information Rights, Policy and Practice*, vol. 1, No. 1, 2016, pp. 3-5.

[135] O'NEIL, C., *Weapons of Math Destruction…*, *cit.*, 2017, p. 85.

[136] OSWALD, M. & GRACE, J., "Intelligence, policing and the use of algorithmic analysis…", *cit.*, 2016, p. 4.

[137] *Ibidem.*

[138] *Ibid.*

The third and final tier of algorithmic decision-making in law enforcement includes systems which evaluate individual risk and behaviour.[139] These tools are used, for instance, in detecting which personal characteristics make a person more likely to break the law and which legal activities may suggest that someone will commit a crime in the near future, for example, the fact that someone purchases a large quantity of plastic bags may indicate a higher probability that he or she will be dealing drugs.[140]

Algorithms encompassed within this third category are also used in recidivism models in order to determine sentence length, parole rights and other elements of convicted criminals' sentences.[141] In order to do so, convicted offenders are handed a test, in which a series of questions regarding past attitudes, general behaviour and other facts about their life are contained. An algorithm analyses answers to the test and combines the conclusions it reaches with regard to said answers with other observations, thereby supposedly determining their likelihood of re-offending.[142]

In Spain, individual predictive algorithms are used for two very specific purposes in law enforcement. The Spanish National Police use VeriPol, a programme that, through the use of machine learning algorithms, detects when an individual has filed a false robbery police report.[143] Another algorithm developed, VioGén, establishes the risk that victims of gender-based violence will suffer further attacks.[144]

---

[139] MIRÓ-LLINARES, F., "Predictive policing: utopia or dystopia? On attitudes towards the use of big data algorithms for law enforcement", *Revista de Internet, Derecho y Política*, No. 30, 2020, pp. 3-5.
[140] FERGUSON, A. G., "Big data and predictive reasonable suspicion", *University of Pennsylvania Law Review*, vol. 163, No. 2, 2015, p. 335.
[141] O'NEIL, C., *Weapons of Math Destruction…*, *cit.*, 2017, pp. 24-27; RITTER. N., "Predicting recidivism risk: new tool in Philadelphia shows great promise", *National Institute of Justice Journal*, No. 271, 2013, pp. 4-13; DRESSEL, J. & FARID, H., "The accuracy, fairness, and limits of predicting recidivism", *Science Advances*, vol. 4, No. 1, 2018.
[142] O'NEIL, C., *Weapons of Math Destruction…*, *cit.*, 2017, pp. 24-27.
[143] QUIJANO-SÁNCHEZ, L. *et al.*, "Applying automatic text-based detection of deceptive language to police reports: extracting behavioral patterns from a multi-step classification model to understand how we lie to the police", *Knowledge-Based Systems*, vol. 149, 2018, pp. 155-168; KOLOTÚSHKINA, N., "VERIPOL: la herramienta de la Policía para detectar denuncias falsas", *RTVE*, 2nd November 2018. Available on 19th February 2019 at: http://www.rtve.es/
[144] CABALLÉ-PÉREZ, M. *et al.*, "El quebrantamiento de las órdenes de protección en violencia de género: análisis de los indicadores de riesgo mediante el formulario vpr4.0", *Anuario de Psicología Jurídica*, No. 30, 2020, pp. 63-72; DEL CASTILLO, C., "Contra la violencia machista, el odio y las denuncias falsas: los algoritmos que usa la Policía", *eldiario.es*, 1st January 2019. Available on 11th April 2019 at: https://www.eldiario.es/; GONZÁLEZ ALVÁREZ, J. L., "Sistema de seguimiento integral en los casos de violencia de género (sistema viogén)", *Cuadernos de la Guardia Civil: Revista de Seguridad Pública*, No. 56, 2018, pp. 83-102.

3.2.2.2. Other algorithmic applications in the exercise of administrative regulatory and coercive powers

Tax authorities also collect and have available very large amounts of personal data and generally have the power to request additional information from public and private entities. For example, in Spain, profiles are drawn in order to detect taxation fraud. Said profiles contain information provided by energy companies and banks as well as data extracted from online rental platforms and from other sources that allow the Spanish Taxation Agency to elaborate profiles that include information on citizens' behaviour as well as their relationships with other natural and legal persons.[145]

Another form of predicting illicit activity is the one developed by algorithmic systems which are oriented to the detection of specific types of behaviours. For example, in Spain, the region of Valencia passed an Act[146] in November of 2018 for the implementation of an automated system for the prevention of bad practices in public administration bodies. The main objective of this automated system is the early detection of corrupt practices in order to prevent the actual commission of criminal offences.[147]

Algorithms can also be used in order to fight discrimination. For example, the Ministry of Interior in Spain is developing an algorithm that detects hate speech, although the objective is not as much felony detection as it is detecting where insults and violence in social media originate.[148]

4. TYPES OF ALGORITHMIC DECISION-MAKING

The scholarship has developed a series of different classifications depending on the way in which algorithms work and the purposes for which they are used. The following pages offer a brief overview of these classifications.

---

[145] VIÑAS COLL, J., "Así son los superordenadores de Montoro contra el fraude fiscal", *Cinco Días*, 24th July 2015. Available on 16th July 2019 at: https://cincodias.elpais.com/; LÓPEZ ZAFRA, J. M., "Patrones de comportamiento y voracidad fiscal", *El Confidencial*, 14th July 2018. Available on 16th July 2019 at: https://blogs.elconfidencial.com/

[146] Act 22/2018 of the Valencian government on the general inspection of services and on the system of alerts for the prevention of bad practices in the Valencian public administration and its instrumental public sector.

[147] CAPDEFERRO VILLAGRASA, O., "El análisis de riesgos como mecanismo central de un sistema efectivo de prevención de la corrupción. En particular, el sistema de alertas para la prevención de la corrupción basado en inteligencia artificial", *Revista Internacional de Transparencia e Integridad*, No. 6, 2018, pp. 1-7.

[148] *Ibidem.*

4.1. AUTOMATIC AND AUTONOMOUS SYSTEMS

Algorithms can be divided in two categories according to their degree of autonomy. YEUNG labels said categories as systems with functional and decisional autonomy.[149] However, for the purposes of this research, it is more useful to differentiate between automatic systems and systems with decisional autonomy.

Automatic systems are those which are able to carry out the purposes they are designed for with no human intervention and which will provide the given outcome if the necessary elements concur. These systems are designed to follow simple instructions and have little or no margin for interpretation. For example, an automatic system deployed to determine whether an individual is eligible for some form of welfare aid may be designed to simply review whether the boxes ticked by the applicant in the relevant form make her eligible for aid. In this case, the algorithm does not analyse further information and does not control that the information contained in the application is true or false, but simply processes the boxes that were ticked and yields an outcome that is already predetermined.

Within the context of administrative law, using automatic systems would be equivalent to the exercise on non-discretionary powers. For example, in the case of the Spanish energy bond, the system had to simply check that applicants complied with the requirements set by the relevant regulatory instrument. Since these systems are simply supposed to act in an automatic manner as they have been instructed they should, in theory, not lead to many errors and other problematic issues that are more frequent in systems with decisional autonomy. However, as it was previously pointed out with regard to the Spanish energy bond example, these systems can also yield erroneous outcomes, for instance, by denying public aid benefits to individuals that comply with all the necessary requirements.

Systems with decisional autonomy (autonomous systems) are those that constantly self-develop and the outcomes of which are not completely predetermined when they are designed. These systems pose greater problems when used both in the private and public sectors, although especially in the latter, due to the loss in control and legitimacy of the decision-making process.

[149] YEUNG, K., "Why worry about decision-making by machine?", in YEUNG, K. & LODGE, M., (Eds.), *Algorithmic regulation*, Oxford, Oxford University Press, 2019, p. 22.

4.2. AUTOMATED AND SEMI-AUTOMATED SYSTEMS

Another classification is the one that differentiates between fully automated and semi-automated decision-making systems. While in the former there is no significant human intervention, the latter does involve some degree of human involvement and oversight, meaning the individual or group in charge of controlling the algorithm can overturn the decisions it makes. Semi-automated decision-making systems can also be used as tools for support in human decision-making processes. This classification is particularly relevant given the fact that more legal constraints are generally placed on fully automated decision-making systems.[150] It is therefore important to analyse the differences in the risks that automated and semi-automated systems may pose on individuals for, as will be discussed later on, the assumption made by much of the scholarship and regulators that many of the risks generated by fully automated systems can be solved by introducing a human-in-the-loop is not necessarily accurate given that humans tend to be biased towards accepting the recommendations made by automated systems.[151]

4.3. PROFILING AND AUTOMATED DECISION-MAKING: DESCRIPTIVE, PREDICTIVE, CLASSIFICATION AND RECOMMENDATION PURPOSES

It is important to highlight the role profiling currently plays in algorithmic decision-making for two main reasons. On the one hand, article 22 of the GDPR specifically mentions profiling as a form of automated processing when it indicates that "the data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling which produces legal effects concerning him or her or similarly significantly affects him or her", which highlights the relevance of this form of data processing.

On the other hand, and more importantly, profiling is the key tool through which algorithms are used in decision-making processes that affect humans. Profiling is generally used as the first step in the decision-making process. Profiles are used in order to predict an individual's behaviour and are therefore the basis upon which the algorithm will produce

[150] For example, article 22 of the GDPR prohibits decisions based solely on automated processing.
[151] CITRON, D. K., "Technological due process", *Washington University Law Review*, vol. 85, No. 6, 2008, pp. 1271-1272; PARASURAMAN, R. & MILLER, C. A., "Trust and etiquette in high-criticality automated systems", *Communications of the ACM*, vol. 47, No. 4, 2004, p. 52.

recommendations or automated decisions.[152] Moreover, the creation of profiles is, in itself, a form of automated decision-making for the profiling algorithm makes decisions regarding the categories in which the individuals whose data is processed will be classified into and the parameters that will be measured and evaluated.

Within profiling, algorithms can be used for (1) descriptive and (2) classification or predictive purposes, whereas algorithms used in automated decision-making are used for (3) recommendation purposes.[153] These three objectives tend to work together.

Algorithms used for descriptive purposes establish patterns and relationships between different pieces of data. This is especially relevant towards creating profiles. For example, an algorithm used by taxation authorities might determine that there is a correlation between making large deductible donations and tax evasion. Classification or predictive systems are the next step in the process. They work by establishing a series of categories or classes that indicate that if a series of data points concur in an individual she will be classified into a certain category.[154] Certain types of behaviour are predicted for all the individuals included in a class or category. Hence, the fact that an individual has made a larger deductible donation, in combination with other pieces of data, will lead to that person being placed in the category that predicts individuals to be at high-risk of committing tax fraud.

Finally, once the descriptive and classification/predictive objectives have been achieved, algorithms can also be used as "systems of recommendation",[155] seeing as once an individual has been predicted to carry out a certain behaviour the machine can recommend what will the best action to address said conduct. This classification by objectives is obviously not as structured or systematic when automated systems operate but it helps to provide an overview of how they work and are used. Algorithms can be used as systems of recommendation in order to inform final decisions made by humans (semi-automated decision-making) or can be directly responsible for making the final decision (automated decision-making).

---

[152] WIEDEMAN, K., "Automated processing of personal data for the evaluation of personality traits: legal and ethical issues", *Max Planck Institute for Innovation and Competition Research Paper No. 18-04*, 2018, p. 15. Available on 29th July 2019 at: https://ssrn.com/

[153] SCHERMER, B. W., "The limits of privacy in automated profiling and data mining", *Computer Law & Security Review*, vol. 27, No. 1, 2011, p. 46; YEUNG, K. & LODGE, M, "Algorithmic regulation: an introduction", in YEUNG, K. & LODGE, M., (Eds.), *Algorithmic regulation*, Oxford, Oxford University Press, 2019, p. 10.

[154] SCHERMER, B. W., "The limits of privacy in automated profiling and data mining", *cit.*, 2011, p. 46; YEUNG, K. & LODGE, M, "Algorithmic regulation…", *cit.*, 2019, p. 10.

[155] YEUNG, K. & LODGE, M, "Algorithmic regulation…", *cit.*, 2019, p. 10.

5. ALGORITHMIC RISKS AND HARMS: GENERAL OVERVIEW

As it was already stated in the introduction, the main phenomenon analysed in this dissertation, that is, algorithmic discrimination, must necessarily be examined with regard to the other risks and harms produced by the increasing use of algorithmic systems. Although this part focuses on addressing algorithmic discrimination within the context of the theoretical and regulatory framework of protection of the rights to equality and non-discrimination, it is useful to contextualise algorithmic discrimination and the perpetuation of inequality through automation in regard to other risks and harms caused by automated systems in order to fully grasp the relevance and implications of the research topic. This section does so in a very brief manner as the dissertation's second part will be dedicated to addressing these issues to a greater extent.

Although there are studies that have shown algorithmic systems to be more efficient and accurate than human decision-makers,[156] they still make mistakes. The amount of data processed by automated systems allows them to reach results based on correlations rather than on causal relationships while still drawing very accurate inferences. Nonetheless, what is true for the group may not be true for an individual on who an inaccurate inference may be made due to the fact that her data points coincide with those of other individuals that have carried out certain actions. For example, an individuals' smart watch may contain information indicating she walks under one kilometre every day. If an algorithm used by a healthcare provider has learned that people who walk under one kilometre every day are generally at greater risk of contracting cardiovascular diseases, it will recommend that she is charged a high insurance prime. However, this particular individual may only use her smart watch when she is not exercising because her purpose is not to control her physical activity but to use other of the watch's applications.[157]

Algorithms can also make mistakes because they are fed biased datasets or because they simply do not work properly because a programmer made a mistake.[158] There is a wide array of reasons due to which an automated system can yield biased or erroneous results. These systems must therefore be controlled and supervised, especially when their results influence

[156] KLEINBERG, J. *et al*., "Human decisions and machine predictions", *The Quarterly Journal of Economics,* vol. 133, No. 1, 2017, pp. 237-293.
[157] MITTELSTADT, B. D. *et al.*, "The ethics of algorithms: mapping the debate", *Big Data & Society*, July-December 2016, p. 6.
[158] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *California Law Review*, vol. 104, No. 3, 2016, pp. 681, 684.

decision-making processes that have significant impacts on the lives of individuals or on society as a whole.

In order for individuals to challenge the automated decisions that they may consider unfair or wrong, it is necessary for them to know the reasons that underlie the algorithmic decisions that affect them and for "technological due process rights"[159] to be set up. With regard to the first issue, recent technological developments have allowed algorithmic systems to become increasingly complex and thus opaque. This complexity hinders the possibility of understanding the underlying logic behind automated decisions unless a detailed and well-crafted explanation is provided. Moreover, in other cases, it is not system complexity, but refusal on the part of processors and controllers to provide information regarding the way in which the system works.[160]

A closely related problem to system opacity, and which hinders both the possibility of rendering automated systems accountable and of providing individuals with due process rights, is opacity in the use of automated systems, particularly in the public sector. The increased efficiency that automating decision-making processes brings about, means that, in many cases, public and private organisations are implementing these systems without providing information regarding the way in which decisions are being made. Thus opacity does not solely apply to the way in which systems work but to general awareness of the extent to which they are being used and for which purposes.[161] Opacity in the automation of public processes is especially problematic given that the transparency principle must necessarily underlie the actions of public powers, except in very specific cases, in order to ensure democratic legitimacy, citizen empowerment and the protection of due process and other fundamental rights. Additionally, the complexity of automated systems can prevent algorithms used by public powers from complying with the justification requirements that that public decision-making must meet.[162]

The effective enactment of due process rights is also heavily hindered by the way in which automated systems are being implemented. Data processing technologies are helping

[159] CITRON, D. K., "Technological due process", *cit.*, 2008, pp. 1249-1313.
[160] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *Fordham Law Review*, vol. 87, No. 3, 2018, pp. 1117-1119.
[161] CATH, C. *et al.*, "Artificial intelligence and the 'good society': the US, EU and UK approach", *Science and Engineering Ethics*, vol. 24, No. 2, 2018, p. 506.
[162] KAMINSKI, M. E., "Binary governance: Lessons from the GDPR's approach to algorithmic accountability", *Southern California Law Review*, vol. 92, No. 6, 2019b, pp. 1545-1549.

organisations to make decisions that affect individuals' lives without structuring systems that allow them to challenge the outcomes produced by algorithms. In this regard, the transparency problem further hampers individuals' chances of contesting algorithmic decisions. In addition, it is not always possible to detect who is responsible for the system, in particular when open source algorithms are used.

It is also essential to point out that, even when there are humans supervising an algorithmic decision their intervention is not always effective. It is necessary for human supervisors to receive proper training on the way the algorithms they will be supervising work. In addition, these humans should also be experts in the area in which the algorithm is being used and have access to the information analysed by the algorithm and relevant information that may not be possible to translate for the automated system. For instance, an automated system used in social services should be supervised by social workers. Humans cannot exercise effective control over a system they do not properly understand, especially if they have no experience in the types of decisions the system is making, particularly if we consider that there is a human tendency to believe machine outputs are accurate and trustworthy.[163]

Algorithms also affect a series of elements that are inherent to human dignity. Automated systems, and the organisations that employ them do not treat individuals as human beings but as measurable objects or products. The amount of digitally stored data on individuals allows for the creation of data doubles upon which predictions on human behaviour are made. The real humans that these data doubles represent, generally have no power over how their digital personality is constructed, thereby losing their agency and autonomy over what part of their personal information is made available to others and leading to harms to the fundamental rights to data protection and privacy.[164]

Finally, algorithms have been shown to produce discriminatory results that affect the members of groups that have been historically disadvantaged in the construction of social power structures (women, racial and religious minorities, etc.) and to help perpetuate the situations of inequality suffered by members of these groups. This problem is closely related with all of the issues detailed above. For example, in order for a victim of discrimination to challenge the decision that harms her rights to equality and non-discrimination, in many cases

[163] DE-ARTEAGA, M. FLOGLIATO, R. & CHOULDECHOVA, A., "A case for humans-in-the-loop: decisions in the presence of erroneous algorithmic scores", in *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, Association for computing machinery, 2020, pp. 1-12.
[164] KAMINSKI, M. E., "Binary governance…", *cit*, 2019b, pp. 1544-1545.

it will be necessary for her to, at least, acquire a minimum understanding of the decision's underlying logic. Additionally, in order to prevent and deal with algorithmic discrimination and the other concerns addressed above, effective oversight mechanisms that comprehensively address as many of the risks generated by algorithms, must be put in place. However, algorithmic discrimination can and should also be tackled by the specific legal framework focused on the protection of the rights to equality and non-discrimination. The following chapters in this part focus on examining the theoretical foundations for the legal protection of equality and on analysing how algorithms discriminate and perpetuate situations of inequality and the way in which the legal framework for the protection of the rights to equality and non-discrimination can address said instances of discrimination.

CHAPTER II. THE THEORETICAL FRAMEWORK TO THE PROTECTION OF EQUALITY AND NON-DISCRIMINATION

This chapter examines the way in which the rights to equality and non-discrimination are constructed and the role that public and private actors play in protecting and respecting said rights. It is important to keep in mind that, while the rights to equality and non-discrimination operate in a wide variety of contexts, this dissertation mainly focuses on the protection of members of disadvantaged groups.

This chapter starts by explaining the nature of fundamental rights. It then offers a general overview of the different theoretical and practical frameworks for the protection of equality and non-discrimination, addressing different concepts of equality and discrimination. The final sections of this chapter are dedicated to explaining the trade-offs that can take place between equality, efficiency and freedom, which is the basis upon which all regulatory and policy instruments that, either directly or indirectly, aim to prevent and deal with algorithmic discrimination, should be based on. The main objective of analysing the trade-off between equality and other competing rights, interests or values is to justify the need for regulations and policies that establish prohibitions to discriminate that apply to both the public and private sector and of measures that ensure the promotion of equality. Said justification is carried out through the proportionality test that has been widely sanctioned and used by European courts.

This chapter aims to establish the theoretical and, to a certain extent, methodological, framework, upon which this dissertation is developed. In this sense, this chapter largely focuses on the need to introduce structural discrimination as a category of analysis when carrying out any proportionality test that involves the rights to equality and non-discrimination of especially protected groups. Since this chapter sets the equality and non-discrimination framework, it will not consider specific rules or policies nor the way in which they should apply to algorithmic discrimination, but generally address the justification for constructing equality and non-discrimination rules and policies from the perspective of substantive equality.

1. THE ROLE, NATURE AND APPLICABILITY OF FUNDAMENTAL RIGHTS

Automated systems are mostly developed by private parties. It is therefore necessary to determine the extent to which these private firms have to build systems that respect the fundamental rights of the individuals whose data they will process and whether the organisations deploying algorithms should respect due process related rights by creating systems that offer the possibility of seeking remedies to unfair decisions. The question that must therefore be answered is whether the protection duties that originate from the recognition of fundamental rights are also applied to private parties.[165]

The *Drittwirkung* theory establishes that fundamental rights must be respected both in their non-interference and protective dimensions not only in the relations between state and private parties but also when private organisations and/or individuals interact with one another.[166] While this theory is generally accepted in the European scholarship, thereby overcoming the initial conceptions of the liberal state in which the fundamental rights of individuals were mainly structured as subjective rights built as mechanisms of protection in their relations with public institutions,[167] there is still no common agreement on whether fundamental rights should apply in a direct or indirect manner to private affairs.

The following pages focus on establishing the nature and role of fundamental rights in order to provide a general overview of the obligation that private parties have in respecting the rights to equality and non-discrimination and the positive obligations that public powers have in protecting said rights.

1.1. FUNDAMENTAL RIGHTS AS PRINCIPLES

The conception of fundamental rights as principles provides them with an overarching nature that exceeds their characterisation as subjective rights that individuals have before public powers. Vesting fundamental rights with the nature of principles means that they act as binding objective rules that underlie the whole legal system. Closely related to this notion is

[165] AGUILERA RULL, A., *Contratación y Diferencia: La Prohibición de Discriminación por Sexo y Origen Étnico en el Acceso a Bienes y Servicios*, València, Tirant lo Blanch, 2013, pp. 31-65; DOMÉNECH PASCUAL, G., *Derechos Fundamentales y Riesgos Tecnológicos*, *cit.*, 2006, pp. 130-134.
[166] BILBAO UBILLOS, J. M., *La Eficacia de los Derechos Fundamentales frente a Particulares: Análisis de la Jurisprudencia del Tribunal Constitucional*, Madrid, Centro de Estudios Políticos y Constitucionales, 1997, pp. 270-276.
[167] *Idem*, pp. 233-240.

the idea that fundamental rights act as binding basic rules in all forms of legal relationships, including private ones, and not just those between public powers and citizens.[168]

1.2. THE DIRECT HORIZONTAL EFFECT OF FUNDAMENTAL RIGHTS

The recognition of the direct horizontal effect of fundamental rights means that said rights directly apply and must be respected in the interactions between private parties. These theories have been criticised for limiting individuals' autonomy and freedom, in particular when the duty to respect the rights to equality and non-discrimination limit the freedom to conduct a business.[169] However, since all fundamental rights have limits, to ensure that all citizens can effectively exercise their freedom, it is sometimes necessary to limit the freedom of others. In this vein, recognising the direct horizontal effect of fundamental rights and thus the possibility that courts have in limiting the freedom of natural and legal persons simply means acknowledging that, in some cases, competing rights will have to be weighed, and that, if deemed adequate, necessary and proportionate, curtailing the freedom of a private party is justified.[170]

Additionally, even though some sectors of the European legal scholarship are still reluctant to accept the direct horizontal effect of the rights to equality and non-discrimination, they do generally concede the *Drittwirkung* of said rights when dealing with unequal treatment cases which result from relationships or interactions between private parties between which there is "a clear imbalance in power".[171]

1.3. THE PROTECTION AND INDIRECT HORIZONTAL EFFECT OF FUNDAMENTAL RIGHTS

1.3.1. Rights to protection

Stricter conceptions of negative liberty in individuals' relations with the state have been long overcome by the recognition of public powers' duty to actively protect fundamental rights.[172] The recognition of rights to protection can be, to a certain extent, constructed and explained in opposition to public non-interference duties in the private sphere of individuals.[173] The distinction between the state's positive and negative obligations is not always relevant. For

[168] AGUILERA RULL, A., *Contratación y Diferencia…*, *cit.*, 2013, pp. 32-33.
[169] *Idem*, p. 36.
[170] DOMÉNECH PASCUAL, G., *Derechos Fundamentales y Riesgos Tecnológicos*, *cit.*, 2006, p. 133.
[171] GERARDS, J., *Judicial Review in Equal Treatment Cases*, Leiden, Koninklijke Brill NV, 2005, p. 27.
[172] ALEXY, R. O., "On constitutional rights to protection", *Legisprudence*, vol. 3, No. 1, 2009, p. 2.
[173] DOMÉNECH PASCUAL, G., *Derechos Fundamentales y Riesgos Tecnológicos*, *cit.*, 2006, pp. 69-72.

example, the European Court of Human Rights' (ECHR) uses very similar criteria to determine failure to comply with positive and with negative state obligations.[174]

Nonetheless, it is useful to differentiate between active and non-interference obligations to the extent that judicial remedies are mainly designed to redress the harms caused by the intervention of public institutions in the sphere of individuals' fundamental rights. Conversely, and while the possibility of claiming that the state did not comply with its active obligations can, in some cases, be successfully brought before courts and other public bodies,[175] it is much harder to prove and obtain remedies for states' failure to comply with their active duties to protect fundamental rights.[176] The limited scope for redress in cases of public powers' omission to actively protect fundamental rights is also indicative of the persisting fear of excessive state protectionism and, sometimes, of the failure to recognise that non-interference can be, in itself, a form of interference when it helps perpetuate the power imbalances that lead to the systematic disadvantage of certain social groups.

While the discussion regarding the notion of fundamental rights as rights to protection and its criticisms can be extensively developed, what is relevant for the purposes of this research is to acknowledge that European case law recognises the existence of states' positive obligations to protect fundamental rights, an idea with which most of the scholarship also agrees.[177] The existence of a state obligation to actively protect fundamental rights is necessary in order for the effectiveness of said rights: if the fundamental rights of citizens and, in particular, of vulnerable members of the community, are not protected, powerful social and economic actors would have the power to coerce and limit the freedom of anyone that is not in a similar position to theirs.[178]

---

[174] *Idem*, pp. 71-72.

[175] European Committee of Social Rights Decision 11th September 2013, Complaint No. 81/2012, European Action of the Disabled (AEH) v. France, paragraph 115: "The Committee notes that there is strong evidence to indicate that France is not fulfilling its obligation, under Article 15§1, to ensure that, in the context of care provision for children and adolescents suffering from autism within specialised institutions such as IMEs or day-hospital units, the work done by these institutions and the working methods they utilise are predominantly of an educational nature".

[176] DOMÉNECH PASCUAL, G., *Derechos Fundamentales y Riesgos Tecnológicos*, *cit.*, 2006, p. 72; ECHR Judgments 21st February 1990, 9310/81, Powell and Rayner v. The United Kingdom, paragraph 41; 9th December 1994, 16798/90, López Ostra v. Spain, paragraph 51 and 12th June 2003, 35968/97, Van Kück v. Germany, paragraph 71.

[177] DOMÉNECH PASCUAL, G., *Derechos Fundamentales y Riesgos Tecnológicos*, *cit.*, 2006, p. 113.

[178] POPPER, K., *The Open Society and its Enemies: Volumes I and II*, Princeton, Princeton University Press, 5th ed., 1962, p. 323: "…*the freedom paradox*. Freedom, we have seen, defeats itself, if it is unlimited. Unlimited freedom means that a strong man is free to bully one who is weak and to rob him of his freedom. This is why we demand that the state should limit freedom to a certain extent, so that everyone's freedom is protected by law.

1.3.2. The indirect horizontal effect of fundamental rights

Theories that defend the indirect horizontal applicability of fundamental rights to private affairs are widely accepted in Germany and contend that fundamental rights bind private parties to the extent that said binding character is contained in legal instruments.[179] The non-essential content of fundamental rights can vary depending on the evolution of society; it is therefore necessary for said content to be determined through legal instruments that respond to the context in which they are created and which are more easily modifiable than Constitutions and international texts that recognise fundamental rights.[180] However, what proponents of the direct horizontal effect theory contend is not that parliaments must develop the content of said fundamental rights that are already recognised in constitutional texts and are therefore binding for all citizens, but that for those rights to be directly applicable in private relations, it is necessary for a legal instrument to recognise them.[181] Additionally, this theory does also admit, to a certain extent, that courts take into consideration the fundamental rights recognised in constitutional texts when interpreting the legal instruments that are applicable to each particular case.[182]

Only recognising the "irradiating" effect of fundamental rights in the framework of the legal instruments that develop them, deprives them of their nature as constitutional rights that can be invoked and applied even when not contained in a legislative act or when a legislative act contains a provision that contradicts them.[183] Nonetheless, the recognition of a series of positive obligations that states have in the protection of fundamental rights enables the redirection of claims regarding harms caused to fundamental rights by private parties. This logic has been applied by the European Court of Human Rights in several cases as it has considered that, by failing to enact their positive obligations to prevent and deal with instances of harms to the rights contained in the European Convention of Human Rights that take place between private parties, states are responsible for the harms caused on said rights

---

Nobody should be at the *mercy* of others, but all should have a *right* to be protected by the state." (Italics in original text).

[179] BILBAO UBILLOS, J. M., "La consolidación dogmática y jurisprudencial de la *Drittwirkung*: una visión de conjunto", *Anuario de la Facultad de Derecho de la Universidad Autónoma de Madrid*, No. 21, 2017, pp. 52-53; DOMÉNECH PASCUAL, G., *Derechos Fundamentales y Riesgos Tecnológicos*, *cit.*, 2006, pp. 130-131

[180] BILBAO UBILLOS, J. M., "La consolidación dogmática y jurisprudencial de la *Drittwirkung*…", *cit.*, 2017, p. 54.

[181] AGUILERA RULL, A., *Contratación y Diferencia…*, *cit.*, 2013, pp. 40-42.

[182] BILBAO UBILLOS, J. M., "La consolidación dogmática y jurisprudencial de la *Drittwirkung*…", *cit.*, 2017, pp. 56-57.

[183] BILBAO UBILLOS, J. M., "La consolidación dogmática y jurisprudencial de la *Drittwirkung*…", *cit.*, 2017, pp. 54-55; DOMÉNECH PASCUAL, G., *Derechos Fundamentales y Riesgos Tecnológicos*, *cit.*, 2006, p. 131.

as a result of private interactions.[184] However, the theory of positive obligations also has its limits as the extent to which states have the responsibility to prevent and act in the face of harms caused to fundamental rights by private parties is limited and depends on the context in which said harm takes place.[185]

2. THE PROTECTION OF THE FUNDAMENTAL RIGHTS TO EQUALITY AND NON-DISCRIMINATION: GENERAL ANALYSIS

2.1. ALGORITHMIC DISCRIMINATION

From a legal perspective, discrimination can be defined as the action of treating a physical or legal person or group of people in a manner that is worse than the way in which other or others in a comparable situation are treated. Discrimination can therefore occur in a wide variety of contexts and situations. However, there are a series of discriminatory actions that are considered especially harmful by democratic states and international human rights instruments. These are instances of discrimination that occur based on categories or grounds that are *a priori* "suspect", such as race or sex. The rationale behind establishing these especially protected categories is that it is unacceptable to base decisions on grounds (characteristics) that are, in principle, immutable (such as race or sex) or that belong to the sphere of autonomy of the individual (such as religion and political opinions), especially when said characteristics are not relevant for the purposes for which the decision is made.[186]

Within each protected category there are certain sub-categories that identify the members of groups that are considered especially vulnerable and/or have historically suffered oppression and disadvantage, such as non-white populations, women or individuals that come from lower socioeconomic backgrounds. Members of these groups are still subjected to discriminatory treatment as a result of stereotypes and prejudices held against them and of social norms and institutions having been built from the perspective of and for those who have traditionally held positions of power. Hence, the special protection offered to instances of discrimination based on suspect categories is also largely aimed to protect these disadvantaged groups. That is, it is obviously inherently wrong to discriminate an individual because he is a man. However, due to the influence of historical power constructions in

[184] ARZOZ SANTIESTEBAN, X., "La eficacia del CEDH en las relaciones entre particulares", *Anuario de la Facultad de Derecho de la Universidad Autónoma de Madrid*, No. 21, 2017, pp. 161-169.
[185] *Idem*, pp. 169-170.
[186] GERARDS, J., "The discrimination grounds of article 14 of the European Convention on Human Rights", *Human Rights Law Review*, vol. 13, No. 1, 2013, p. 114.

society, which will be explained to a greater extent later on in this chapter, it is not men but women that are generally the victims of structural and specific cases of discrimination. This means that when analysing both particular cases and the general discrimination suffered by members of vulnerable or disadvantaged groups, it is necessary to consider the historically oppression and current existence and pervasiveness of structures of discrimination that affect members of said groups.

The specific importance and pervasiveness of the discrimination suffered by members of disadvantaged groups is the main focus of this dissertation, in particular, of this first part. Hence, in order to clarify the terminology used, it is important to note that when references to protected or suspect grounds, categories, attributes or characteristics are made, they will refer to the general characteristic or ground (sex, race, age, etc.). However, when concepts such as "disadvantaged group", "protected group", "especially protected group" and "oppressed group" are used, they aim to encompass only the sub-categories of protected characteristics which are especially vulnerable to both structural and specific instances of discrimination.

In order to differentiate between situations of discrimination in which an individual is treated unfairly based on protected grounds and instances of discrimination that result from a context of domination or disadvantage, BARRÈRE UNZUETA has suggested using the concept of "subordiscrimination" for the latter.[187] While keeping in mind that this dissertation focuses on situations of discrimination mediated by contexts of oppression and disadvantage, both "discrimination" and "subordiscrimination" will be indistinctly used. Therefore, when "discrimination" or any other related concept is used, it generally refers to the unfair and unequal treatment of members of vulnerable and disadvantaged groups.

Instances of unfair treatment undergone by members of vulnerable and disadvantaged groups are generally the result of structures of discrimination that have been developed over centuries and that still persist in our societies, thereby leading to what can be labelled as "institutional discrimination". Discriminatory structures and notions are deeply and, in many cases, inadvertently, embedded at the core of Western democracies. These forms of unconscious yet highly pervasive discrimination lie at the foundation of algorithmic

[187] BARRÈRE UNZUETA, M. A., "Iusfeminismo y derecho antidiscriminatorio: hacia la igualdad por la discriminación", in MESTRE, R., (coord.), *Mujeres, Derechos y Ciudadanías*, Valencia, Tirant Lo Blanch, 2008, pp. 45-72.

discrimination.[188] Since automated systems are created by humans and are based and learn from real-world information, they can easily reproduce certain prejudices and discriminatory structures if not carefully reviewed. In this respect, over the past few years, the scholarship and human rights activists have brought forward multiple examples of automated systems that generate discriminatory outcomes for protected groups and of algorithms that help perpetuate certain negative stereotypes and situations of disadvantage. Thus, generally, when I refer to algorithmic discrimination I refer to specific instances in which algorithms discriminate against members of disadvantaged groups or in which algorithms help to perpetuate structures of inequality and subordination suffered by members of said groups.

Algorithms can also produce unfair results that are not based on individuals' protected group membership. In addition, algorithms generate a series of risks for a number of fundamental rights, such as the rights to privacy and due process, as well as for several democratic principles, such as transparency related mandates. Some of the solutions adopted for said risks are closely related to the hazards that algorithmic systems create with regard to the rights to equality and non-discrimination and will be addressed in the second part of the dissertation. However, since the main focus of this dissertation is algorithmic discrimination and given the particularities of equality and discrimination both as social and legal phenomena, this part of the dissertation focuses on examining how algorithmic discrimination takes place and the way in which the rights to equality and non-discrimination in algorithmic decision-making can be articulated through the anti-discrimination and substantive equality frameworks.

2.2. THE EQUALITY AND ANTI-DISCRIMINATION FRAMEWORK

Legal instruments aimed towards preventing and dealing with instances of discrimination can be grouped into two main categories: *ex ante* and *ex post* mechanisms. Prohibitions to discriminate are *ex post* mechanisms while blindness, equality mainstreaming, promotion of equality measures and affirmative actions are *ex ante* mechanisms.

[188] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, pp. 673-674.

2.2.1. Prohibitions to discriminate and concepts of discrimination

2.2.1.1. Prohibitions to discriminate: direct and indirect discrimination

Anti-discrimination law, which focuses on prohibiting and setting mechanisms to detect and punish different forms of discrimination, can be classified as an *ex post* mechanism. Anti-discrimination law can be categorised as an *ex post* mechanism for, although its nucleus is conformed by prohibitions to discriminate, said prohibitions are mainly enacted as judicial remedies to fight discrimination once it has happened. Hence, while the aim is to dissuade parties to discriminate through the threat of being punished, since said punishment is enacted once the action takes place, it can be seen as an *ex post* mechanism. Nonetheless, as it will be subsequently argued, the two main types of discrimination considered under anti-discrimination law, that is, direct and indirect discrimination, can be conceptually aligned with the two types of mechanisms aimed towards preventing instances of discrimination: anti-classification and anti-subordination tools.[189]

Cases of direct discrimination take place when decisions produce instances of discrimination based on suspect grounds or characteristics. For example, a clear case of direct discrimination is denying access to a restaurant to members of certain ethnic groups. Thus, to a certain extent direct discrimination takes care of protecting formal equality in that it focuses on ensuring that no decisions privilege one group over another on the basis of a protected ground.[190] However, direct discrimination can also occur when measures of positive action that are legally mandated and should be taken in order to grant equal opportunities to disadvantaged groups are not carried out.[191] This particular version of direct discrimination is thus associated with substantive forms of equality.

Indirect discrimination takes place when an apparently objective rule places one sex, race, religion, sexual orientation, etc., at a disadvantage with regard to other race, sex, religion, sexual orientation, etc.[192] Indirect discrimination therefore does not focus on the treatment

[189] BORNSTEIN, S., "Antidiscriminatory algorithms", *Alabama Law Review*, vol. 70, No. 2, 2019, p. 542.
[190] KÜLLMANN, M., "Platform work, algorithmic decision-making, and EU gender equality law", *The International Journal of Comparative Labour Law and Industrial Relations*, vol. 34, 2018, p. 12.
[191] AÑÓN ROIG, M. J., "Grupos sociales vulnerables y derechos humanos. Una perspectiva desde el derecho antidiscriminatorio", in ANSUÁTEGUI ROIG, J., *et al.*, (coords.), *Historia de los derechos fundamentales*, Madrid, Dykinson, 2013a, p. 646.
[192] XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence: mapping the challenges of algorithmic discrimination", in BERNITZ, U. *et al*., (eds.), *General Principles of EU Law and the EU Digital Order*, Alphen aan den Rijn, Kluwer Law International, 2020, p. 170.

provided by these decisions, which is apparently neutral, but on the actual impact they have on the group as a whole.

2.2.1.2. Other forms of discrimination

There are also other forms of discrimination that serve as categories of analysis rather than as legal classifiers for prohibited actions.

i) *Structural discrimination*

Structural (or systemic) discrimination refers to the way in which the dominant narratives that have historically disadvantaged the members of certain groups such as women and non-white populations, which will be analysed later on to a greater extent, are intrinsic to the development of social norms and structures. Structural discrimination results from the construction of social values from the perspective of those who have historically held power and which have come to represent the universal liberal (and autonomous) individual: wealthy white men.[193] The specific materialisations of situations of discrimination result from the existence of general discriminatory structures.

It is important to differentiate between structural discrimination (oppression or subordination) and specific cases of discrimination. The former refers to the construction and portrayal of certain elements that conform individuals' identity through a narrative of subordination. It requires the creation of general categories (such as race), within which a hierarchy of sub-categories is established (for example, white is better than black). These narratives of subordination have been developed throughout history and are therefore embedded at the core of social structures and institutions. Being a member of an oppressed identity-group therefore means being part of a society the structures of which are underpinned by certain principles that discriminate against those classified within said group (structural discrimination).[194]

The construction of social structures and institutions through narratives of oppression leads to specific forms of discrimination. Consequently, structural discrimination and specific instances of direct and indirect discrimination can be analysed as part of the same process or

[193] BARRÈRE UNZUETA, M. A., "Igualdad y 'discriminación positiva': un esbozo de análisis teórico-conceptual", *Cuadernos Electrónicos de Filosofía del Derecho*, No. 9, 2003b, pp. 7-8.
[194] *Ibidem*.

continuum by which, the narratives of oppression that have been built and become a part of society throughout history, lead to specific situations in which members of the hierarchically inferior identity-groups are discriminated against.[195]

ii) *Intersectional discrimination*

In many cases, not one but several of the characteristics that place individuals at risk of being discriminated may be part of an individuals' identity. For example, women tend to live longer than men, to accumulate less resources during their lifetime and to have some form of chronic illness or impairment that is not fatal but that leads to them living during more years with abilities that differ from the norm.[196] The situations of discrimination suffered by these women are further aggravated if they also belong to an ethnic/racial minority.[197] Additionally, the discrimination of older people in the workplace affects women at a much higher rate than it affects men.[198]

In order to analyse the specific forms of discrimination suffered by individuals who fall under different identities and to foster tools that allow for their particularities to be considered and addressed, the concept of intersectionality, sometimes also known as multiple discrimination, was developed.[199] Its origins lie in the black feminist critique to theories of discrimination that focused exclusively on gender or race,[200] and has since developed in order to cover many other identities.

Intersectionality is key because all the different narratives that have historically built society through oppressing and disadvantaging certain groups, heavily rely on treating individuals within said groups as homogeneous, thus further reinforcing negative stereotypes and the lack

[195] AÑÓN ROIG, M. J., "Principio antidiscriminatorio y determinación de la desventaja", *Isonomía: Revista de Teoría y Filosofía del Derecho*, No. 39, 2013b, pp. 127-157.
[196] BERRIDGE, C. W. & MARTINSON, M., "Valuing old age without leveraging ableism", *Generations,* vol. 41, No. 4, 2018, p. 85.
[197] GONZÁLEZ RAMS, P., "Las mujeres con discapacidad y sus múltiples desigualdades; un colectivo todavía invisibilizado en los estados latinoamericanos y en las agencias de cooperación internacional", in REY TRISTÁN, E. & CALVO GONZÁLEZ, P., (coords.), *200 Años de Iberoamérica (1810-2010)***,** 2010, pp. 2737-2756.
[198] NEUMARK, D., BURN, I. & BUTTON, P "Is it harder for older workers to find jobs? "Is it harder for older workers to find jobs? New and improved evidence from a field experiment", *Journal of Political Economy*, vol. 127, No. 2, 2019, p. 966.
[199] GONZÁLEZ RAMS, P., "Las mujeres con discapacidad y sus múltiples desigualdades…", *cit*., 2010, pp. 2747-2748.
[200] CRENSHAW, K., "Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics", *University of Chicago Legal Forum*, No. 1, 1989, p. 140.

of attention paid to their particularities as part of the group but also as individuals.[201] This is especially relevant when an individual belongs to more than one identity-group, something that had gone largely ignored by research on discrimination, as it tended to single out and focus one of the axes, such as gender or race.[202]

The need to address situations of discrimination from more than one axis of oppression results from the fact that individuals whose identity crosses more than one disadvantaged group endure worse conditions.[203] However, this need also follows from the fact that these identities do not overlap with one another but create specific narratives applicable say, for example, to black women.[204]

The main problem with intersectional approaches to discrimination is the fact that they do not offer a closed list of categories to analyse. As BUTLER puts it:

> "Theories of feminist identity that elaborate predicates of colour, sexuality, ethnicity, class and able-bodiedness invariably close with an embarrassed 'etc.' at the end of the list. Through this horizontal trajectory of adjectives, these positions strive to encompass a situated subject, but invariably fail to be complete".[205]

While the categorisations resulting from dominant narratives are rejected in critiques to said discourses, social reality, and thus the oppression that said categories produce, can not be neglected.[206] Consequently, while the discrimination suffered by individuals must be carried out from intersectional approaches by analysing the different narratives affecting each particular case, it is very difficult to encompass all discrimination axes in all research pieces. Nevertheless, when developing works based on theories of liberation, the limits (and necessity) of addressing exclusively some but not all situations of oppression must be acknowledged. Additionally, when individual cases of discrimination are examined, for instance, when claims are brought before courts, it is essential for the intersectional perspective to be included in the analysis.

---

[201] CELIOUS, A. & OYSERMAN, D., "Race from the inside: an emerging heterogeneous race model", *Journal of Social Issues*, Vol. 57, No. 1, 2001, p. 151; LUDVIG, A., "Differences between women? Intersecting voices in a female narrative", *European Journal of Women's Studies*, vol. 13, No. 3, 2006, p. 246.
[202] LUDVIG, A., "Differences between women?...", *cit.* 2006, p. 246.
[203] NOBLE, S. U., *Algorithms of Oppression*: *How Search Engines Reinforce Racism*, New York, New York University Press, 2018, pp. 93-94.
[204] *Idem*, pp. 94-96.
[205] BUTLER, J., *Gender Trouble*, London, Routledge, 1990, p. 143.
[206] LUDVIG, A., "Differences between women?...", *cit.* 2006, p. 248.

iii) *Discrimination by indifferentiation*

Discrimination by indifferentiation is not just useful as a category of analysis but is also a specific type of discrimination. Discrimination by indifferentiation takes place when cases that are substantially different are treated the same.[207]

2.2.2. Anti-classification, anti-subordination and concepts of equality

There are two types of *ex ante* mechanisms aimed towards preventing discrimination: anti-subordination and anti-classification mechanisms. This specific classification is particularly important in the US and has therefore been mostly developed by the American legal scholarship.[208] However, given the fact that the GDPR is the main instrument specifically aimed towards preventing instances of discrimination in automated decision-making and that, as a privacy-based instrument, it mainly draws from the idea of anti-classification, it is vital to introduce said categories into the European analysis of algorithmic discrimination.

2.2.2.1. Anti-classification and formal equality

The close relationship between anti-classification and data protection stems from the fact that the former draws from notions of privacy and blindness to protected attributes such as race or sex. This is precisely the idea that articles 9 of the GDPR and 11 of Directive 2016/680 for data protection in law enforcement[209] follow from when they prohibit the processing of "special categories of personal data", seeing as these categories more or less align with the protected categories recognised in international fundamental rights instruments.

Anti-classification strategies, which are especially common in the US, mainly aim towards protecting individuals' privacy instead of proactively shifting dominant narratives.[210] The underlying idea to this type of anti-discrimination regulation and policy is that elements such as an individual's sex or race should not be considered, even if this consideration were to take

[207] SALOMÉ, L. M., "La discriminación y algunos de sus calificativos: directa, indirecta, por indiferenciación, interseccional (o múltiple) y estructural", *Revista de Pensamiento Constitucional*, vol. 22, No. 22, 2017, pp. 267-271.
[208] BALKIN, J. M. & SIEGEL, R. B., "The American civil rights tradition: anticlassification or antisubordination?", *University of Miami Law Review*, vol. 58, No. 1, 2003, pp. 9-34.
[209] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA.
[210] ROBERTS, J. L., "Protecting privacy to prevent discrimination", *William & Mary Law Review*, vol. 56, No. 6, 2015, p. 2123.

place with positive outcomes for members of the traditionally oppressed group.[211] These instruments therefore focus on formal equality.

Formal equality stems from the Aristotelian principle that contends that similar treatment should be provided to similar cases.[212] This form of equality is widely accepted from all liberal perspectives. Formal equality requires compliance with the principle of equality before the law. It is also closely related to the prohibitions of direct discrimination, which require eliminating provisions and behaviours that treat someone less favourably than another individual in a comparable situation, particularly when the difference in treatment is based on certain characteristics, such as sex or race. Hence, anti-classification mechanisms approach discrimination from the perspective of formal equality as it is thought that, by providing the same treatment to all individuals with no consideration towards disadvantaged group membership, it will be possible to achieve real equality of opportunity.[213]

Anti-classification is commonly expressed through the idea of "blindness", which basically entails that the decision-maker treats the decision subject as though her race and sex or any other protected characteristic were invisible.[214] This type of measure can be especially successful when the party at risk of carrying out a discriminatory conduct is really unaware of said characteristics.[215] It is thought that it is not possible to discriminate against what is not known. In this regard, when privacy regulations focus on preventing the access or processing of information regarding protected attributes, the framework is shifted from notions of autonomy and dignity, which are the main justifications for privacy protections, to the centrality of equality and fairness related norms, thus privacy protection can be analysed or interpreted as a form of antidiscrimination regulation.[216]

Direct discrimination can be analysed in relation to anti-classification mechanisms for it prohibits decisions that harm individuals based on suspect grounds and that generally harm members of disadvantaged groups.[217] Therefore, direct discrimination could be prevented by

[211] *Ibidem*.
[212] GERARDS, J., *Judicial review in equal treatment cases*, *cit.*, 2005, pp. 9-10.
[213] BORNSTEIN, S., "Antidiscriminatory algorithms", *cit.*, 2019, p. 541.
[214] ROBERTS, J. L., "Protecting privacy to prevent discrimination", *cit.*, 2015, p. 2123.
[215] GOLDIN, C. & ROUSE, C., "Orchestrating impartiality: the impact of "blind" auditions on female musicians", *The American Economic Review*, vol. 90, No. 4, 2000, pp. 715-741; ROBERTS, J. L., "Protecting privacy to prevent discrimination", *cit.*, 2015, p. 2123-2124.
[216] ROBERTS, J. L., "Protecting privacy to prevent discrimination", *cit.*, 2015, p. 2122.
[217] BORNSTEIN, S., "Antidiscriminatory algorithms", *cit.*, 2019, p. 542.

not considering elements such as sex or race, unless, of course, the element considered was inextricably linked to the protected ground.[218]

However, it is important to keep in mind that the prohibitions of direct discrimination cannot be totally identified with preventing discrimination through privacy. Provisions that prohibit direct discrimination outlaw making decisions based on the protected characteristic while privacy instruments ban any consideration of said characteristic. For example, considering a person's race in order to ensure that the algorithmic system did not discriminate against racial minorities would be allowed under anti-discrimination law because the decision would not be based on the protected characteristic. It would, however, be prohibited under privacy provisions which completely ban the processing of specially protected categories of data. References to anti-classification instruments must be understood to refer to the use of privacy regulatory instruments to prevent discrimination.

2.2.2.2. Anti-subordination and substantive equality

The anti-discrimination mechanisms that are shaped with the objective of reframing dominant narratives fall within the framework of anti-subordination. These tools focus on substantive equality[219] and encompass strategies of positive differential treatment to correct historical situations of subordination as well as other types of policies, aimed towards eliminating the disadvantage of some groups by integrating their particularities when developing general regulatory instruments and policies and introducing diversity strategies.[220] This framework is largely constructed from a perspective that aims to recognise and integrate the reality of a system that structurally discriminates against the members of certain groups.

i) *Dimensions of substantive equality*

Substantive equality is a concept that can be, and is, structured from multiple perspectives. Firstly, it recognises the existence of disadvantaged individuals and groups and aims to fairly

[218] For example, a rule that only allows survivor's benefit to widows and widowers that were married to their partners is considered direct discrimination against homosexuals if they are not allowed to get married. CJEU Judgment 1st April 2008, C-267/06, Tadao Maruko v. Versorgungsanstalt der deutschen Bühnen, paragraph 72: "If the referring court decides that surviving spouses and surviving life partners are in a comparable situation so far as concerns that survivor's benefit, legislation such as that at issue in the main proceedings must, as a consequence, be considered to constitute direct discrimination on grounds of sexual orientation, within the meaning of Articles 1 and 2(2)(a) of Directive 2000/78."

[219] BALKIN, J. M. & SIEGEL, R. B., "The American civil rights tradition…", *cit.*, 2003, pp. 9-10; BORNSTEIN, S., "Antidiscriminatory algorithms", *cit.*, 2019, p. 541.

[220] ROBERTS, J. L., "Protecting privacy to prevent discrimination", *cit.*, 2015, p. 2123.

redistribute material (i.e.: wealth) and intangible (i.e.: decision-making power) goods.[221] That is, it does not solely focus on neutral rules that do not explicitly discriminate, but also acknowledges that there are a series of discriminatory social structures that disadvantage the members of certain groups. The redistributive dimension of substantive equality is for instance expressed through the principles of NUSSBAUM and SEN's[222] capabilities theory in that it requires providing every individual with actual opportunities (capabilities) to freely develop. Formal equality is thus considered insufficient to provide all individuals with real opportunities. The redistributive dimension of substantive equality necessarily includes providing individuals with resources that ensure their participation and integration in the social and political community.[223]

The second dimension of the notion of substantive equality is developed from the perspective of dignity.[224] This dimension is particularly useful to identify situations in which the disadvantage of groups is perpetuated through actions that do not specifically constitute instances of discrimination but that help to reproduce prejudices against members of certain groups, such as the reproduction of negative stereotypes in advertising. Approaching equality from this perspective also offers a series of criteria to balance out the specific interests of different especially protected groups. For example, while religious customs must be respected, limits have to be placed on them when they denigrate other groups.[225] This dimension is closely related to the redistributive dimension addressed in the previous paragraph in that, by stereotyping and stigmatising members of certain groups, their social and economic opportunities are curtailed. In fact, the actions considered by the dignitary perspective to substantive equality can be analysed as a first step in the process of generating and reproducing discriminatory social structures.

Finally, and as a result of its other dimensions, substantive equality does not aim to pursue fair redistribution by forcing members of the disadvantaged group to imitate the roles and values of the dominant group, but to restructure society in order to accommodate the

---

[221] YOUNG, I. M., *Justice and the Politics of Difference*, Princeton, Princeton University Press, 1990, p. 16.
[222] NUSSBAUM, M., *Women and Human Development: The Capabilities Approach*, New York, Cambridge University Press, 2000; SEN, A., *The Idea of Justice*, Cambridge (Massachusetts), The Belkap Press of Harvard University Press, 2009.
[223] FREDMAN, S., *Discrimination Law*, New York, Oxford University Press, 2nd ed., 2011, pp. 31-33.
[224] *Idem*, pp. 28-30.
[225] *Idem*, p. 29.

particular experiences of all groups, hence ensuring their full participation as citizens.[226] In this sense, substantive equality aims to ensure the protection of freedom within diversity.

Substantive equality goes further than formal equality in recognising the existence of a series of structures that disadvantage the members of certain population groups and aims to change and eliminate the system of subordination suffered by disadvantaged groups. This version of equality therefore helps to shape the anti-subordination framework and is intimately connected with the recognition of indirect discrimination.

ii) *Methodological and policy approaches to substantive equality: anti-subordination mechanisms*

The methodological approaches to the materialisation of substantive equality can be categorised as equality of opportunity or equality of results approaches. The former mainly focuses on removing barriers that disadvantage certain groups and implementing positive measures, such as investments specifically directed towards helping disadvantaged groups.

The policy mechanisms through which equality of opportunity can be made effective are mainly equality mainstreaming and promotion of equality measures. Equality mainstreaming requires framing policies and regulatory instruments through the introduction of the perspective and experiences of disadvantaged groups. It means, for instance, reframing social values with the objective of providing equal value to typically male and female professions.

Equality of results is methodologically useful both for detecting instances of discrimination and for redressing them. Firstly, a case of discrimination may be detected by analysing whether two individuals in comparable situations obtain the same results in a given scenario. For example, if there is no equality of results in the evaluation of a man and a woman in comparable or identical professional situations who apply for a job, the lack of equality of results provides us with information that a case of sex discrimination might be taking place.

Secondly, the policy mechanism that can be used in order to make equality of results effective is affirmative action, also known as positive action. Affirmative or positive action policies are sometimes labelled "positive or reverse discrimination", a concept to which negative connotations are generally attached and which is sometimes used in order to identify

[226] *Idem*, pp. 30-31.

positive action mechanisms that are illegal.[227] The equality of results approach, when adopted, must necessarily be integrated with all the dimensions of substantive equality and the policy mechanisms aimed towards achieving equality of opportunity in order to properly reframe the core of social structures of domination.

Equality of results approaches to equality and non-discrimination and affirmative action policies are the most controversial instruments aimed to achieve substantive equality. The implementation of affirmative action policies is always problematic because it requires establishing rules that harm one (structurally advantaged) group while favouring another (structurally disadvantaged) group. However, what is especially difficult, as it will be discussed to a greater extent later on, is determining the exact level of equality that it is aimed for and the specific consequences that affirmative action measures can have. While it is true that discriminatory outcomes are either directly or indirectly the result of the aforementioned structures of disadvantage, demanding absolute group parity or proportional representation can sometimes not achieve its ulterior aims (dismantling systemic discrimination) and even have counterproductive effects for the group it aims to help or negatively affect other legitimate policy aims to a greater extent than it benefits equality.

Indirect discrimination can be linked to anti-subordination mechanisms for it considers the collective element of discrimination as well as the way in which, apparently neutral social norms, can disproportionately affect members of disadvantaged groups.[228] Indirect discrimination can be the result of the structural discrimination suffered by disadvantaged groups or of the negative stereotypes that they can be associated with. In the first case, the apparently neutral practice or criterion, such as requiring a higher degree of education that what is strictly required to develop a job, will benefit dominant groups who have historically had and still have more access to resources. In the second case, decisions made regarding members of disadvantaged groups will be mediated by the negative stereotypes associated to them, resulting in decisions that perpetuate their subordination. For example, if an employer is looking for someone with leadership skills for a project manager position and the employer has internalised the stereotype that women are bad leaders, all women interviewed will very likely automatically be perceived to not show any leadership skill and the person hired will therefore be male.

[227] BARRÈRE UNZUETA, M. A., "Igualdad y 'discriminación positiva'…", *cit.*, 2003b, pp. 20-21.
[228] BORNSTEIN, S., "Antidiscriminatory algorithms", *cit.*, 2019, p. 542.

This dissertation is mainly built upon notions of substantive equality and aims to highlight the need for framing equality policies and rules within the anti-subordination framework of equality and non-discrimination. The following section explains the conflicts that exist between equality and other competing values and aims to justify the need for the direct horizontal effect of the rights to equality and non-discrimination but also, mainly, the need for protecting the rights to equality and non-discrimination of disadvantaged groups from the perspective of anti-subordination both in the public and private sector.

3. JUSTIFYING THE HORIZONTAL APPLICATION OF THE FUNDAMENTAL RIGHTS TO EQUALITY AND NON-DISCRIMINATION: THE EQUALITY, FREEDOM AND EFFICIENCY TRADE-OFF

3.1. BALANCING CONFLICTING RIGHTS AND INTERESTS

When addressing the general regulatory and policy framework set out in order to protect the rights to equality and non-discrimination, it is important to pay attention and respond to the conflict between aspirations of equality and other competing values, specifically efficiency and freedom. The trade-off between equality, efficiency and freedom is especially relevant for the purposes of this research when materialised as a conflict between equality and the freedom to conduct a business in the private sector. In the public sector, this trade-off has mainly been reflected in the conflict between equality and public security, for example, when risk assessment algorithms accurately detect that members of racial minorities are more likely to reoffend than white individuals.[229] Other specific trade-offs between equality and other public interests and policy effectiveness have also appeared in the public sector.[230] In addition, there are cases in which the conflict between equality and other values can present significant problems, for instance, when freedom of speech is articulated in ways that can be considered to denigrate members of disadvantaged groups.

The existence of a trade-off inevitably means that the effective protection of the rights to equality and non-discrimination of certain individuals or groups requires limiting the freedom of others as well as carrying out actions that are not efficient in economic terms, at least in the short run. Said limitations to other competing rights and interests must therefore be

[229] ANGWIN, J. *et al*, "Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks", *Propublica*, 23rd May 2016.
[230] RANCHORDÁS, S. & SCHUURMANS, Y., "Outsourcing the welfare state: the role of private actors in welfare fraud investigations", *European Journal of Comparative Law and Governance*, vol. 7, No. 2, 2020, pp. 5-42.

properly justified through a proportionality analysis in which the conflicting rights and interests are weighed.

The proportionality principle is used by many European courts in order to develop a rational analysis of the weight that should be given to conflicting rights and interests in different scenarios.[231] The proportionality analysis must necessarily be preceded by the justification of the legitimacy of the aim pursued by the measure that is analysed.[232] Once the existence of a legitimate aim has been properly established, it is possible to carry out the proportionality analysis. The proportionality analysis or test, takes place as a three-step process in which the provision, criterion or practice examined must be successfully proven as adequate (suitable) to satisfy the objective it aims to accomplish; necessary in the sense that there is no other measure that can accomplish the same aims in an equally or similarly effective manner while causing less harm to the competing rights or interests; and, proportional (in the strict sense) to the objectives it aims to accomplish.[233]

The last part of the proportionality test requires the careful consideration and balancing of all interests at play and its resolution is what specifically determines the way in which the trade-off plays out. The *strictu sensu* proportionality test determines the exact importance of each of the rights and interests at play and, according to ALEXY is identical to a rule that can be labelled "The rule of balancing",[234] which determines that: "The greater the degree of non-satisfaction of, or detriment to, one principle, the greater must be the importance of satisfying the other".[235]

This section mainly focuses on justifying the balancing of interests in private relations given the contentious character of limiting the freedom of private parties through equality and non-discrimination mandates. However, the proportionality analysis carried down below is also useful when applied to public sector regulations and policies in which equality must be balanced out with competing interests, such as public security.

---

[231] DOMÉNECH PASCUAL, G., *Derechos Fundamentales y Riesgos Tecnológicos*, *cit.*, 2006, p. 160.
[232] AGUILERA RULL, A., *Contratación y Diferencia…*, *cit.*, 2013, p. 83.
[233] ALEXY, R., "Sobre los derechos constitucionales a protección", in GARCÍA MANRIQUE, R., (Ed.), *Derechos Sociales y Ponderación*, Fundación coloquio jurídico europeo, Madrid, (2nd ed), 2009, pp. 56-65.
[234] ALEXY, R., *A Theory of Constitutional Rights*, translation by Julian Rivers, Oxford, Oxford University Press 2002, p. 102; ALEXY, R., "Constitutional rights and proportionality", *Journal for Constitutional Theory and Philosophy of Law*, vol. 22, 2014, p. 54.
[235] *Ibidem.*

3.2. DO PROHIBITIONS TO DISCRIMINATE AND EQUALITY MANDATES AND MEASURES APPLICABLE TO THE PRIVATE SECTOR PURSUE A LEGITIMATE AIM?

The equal treatment principle and the recognition of the rights to equality and non-discrimination are enshrined in international rights instruments and most national constitutions. Equality mandates and prohibitions to discriminate thus pursue a legitimate aim to the extent that they aim to make the aforementioned rights and principle effective. Additionally, not few discriminatory conducts also harm individuals' dignity, which is also an essential protected value in democratic countries. Finally, and more specifically in private contexts such as employment or the access and provision of goods and services, equality mandates and prohibitions to discriminate also aim to protect the freedom and right to employment and the freedom and right to contract of individuals at risk of being discriminated against.[236]

3.3. ARE PROHIBITIONS TO DISCRIMINATE AND EQUALITY MANDATES AND MEASURES APPLICABLE TO THE PRIVATE SECTOR SUITABLE TO SATISFY THE OBJECTIVES THEY AIM TO ACCOMPLISH?

Determining the suitability of prohibitions to discriminate and equality mandates and measures entails determining whether they contribute to eradicate discrimination. In general, countries that have adopted legislative reforms aimed towards eliminating discrimination and promoting equality have reduced inequality,[237] particularly when the regulatory instruments adopted took the form of hard law.[238]

Since the aim of this section is partly to determine the interactions and trade-offs between equality and efficiency it is also relevant to point out that OKUN argues that in certain cases, when the political process intervenes in the market by regulating and, for example, banning discriminatory actions that may be apparently or really efficient for firms, economic behaviour can also be shifted in the long run, changing the social norms surrounding certain

[236] KEREN, H., "We insist! Freedom now: Does contract doctrine have anything constitutional to say?", *Michigan Journal of Race Law*, vol. 11, 2005, pp. 133-193.
[237] Resolution 2111 (2016) on assessing the impact of measures to improve women's political representation of the Council of Europe's Parliamentary Assembly; Morgenroth, T. & Ryan, M. K., "Quotas and affirmative action: understanding group-based outcomes and attitudes", *Social and Personality Psychology Compass*, vol. 12, No. 3, 2018, pp. 1-14; NORDEN, "Gender equality – The nordic way", Copenhaguen, Nordic Council of Ministers, 2010; OKUN, A. M., *Equality and Efficiency Equality and Efficiency: The Big Tradeoff*, Washington DC, Brookings Institution Press, 2015 (1st ed. 1975), p. 77.
[238] RUBIO, A., "La eficacia de la legislación española en materia de igualdad de género", Gênero & Direito, vol. 4, No. 1, 2015, pp. 76-113.

types of decisions.[239] In fact, this shift in the marketplace's behaviour pushed forward by public regulation can also prove that the discriminatory attitudes banned are inefficient. OKUN provides the example of equal employment opportunity, which not only did significantly improve the employment opportunities and conditions of black individuals in the US, but also increased the US' real gross national product.[240]

3.4. ARE PROHIBITIONS TO DISCRIMINATE AND EQUALITY MANDATES AND MEASURES APPLICABLE TO THE PRIVATE SECTOR NECESSARY?

3.4.1. Explaining discrimination from an economic perspective

Since the conflict is largely approached from the perspective of efficiency defined in economic terms, it is necessary to explain and understand how different forms of discrimination can be classified from an economic perspective. The following classification shows how some forms of discrimination are economically efficient. Although this efficiency is the result of society being built upon structures of subordiscrimination, the direct (positive) economic effects that discriminating can sometimes have can not be ignored and must be taken into consideration when establishing regulatory and policy instruments aimed to combat discrimination. In other words, this section aims to convey the need for prohibitions to discriminate and equality measures and mandates given the fact that discrimination still takes place and can sometimes even be considered "rational" in economic terms. The following classification focuses on how the forms of discrimination explained below operate when discriminating against members of disadvantaged groups.

3.4.1.1. Animus-based discrimination

The first form of discrimination is the one that results from the decision-maker's dislike for members of the protected group. This form of discrimination takes place, for example, when an employer has an aversion towards people of a certain ethnic background and therefore refuses to hire said workers.[241] Out of the different types of discrimination this is the only one that is, in principle, inefficient.

[239] OKUN, A. M., *Equality and Efficiency…*, *cit.*, 2015 (1st ed. 1975), pp. 76-77.
[240] *Idem*, p. 77.
[241] STRAHILEVITZ, L., "Privacy versus antidiscrimination", *University of Chicago Law Review*, vol. 75, No.1, 2008, p. 365.

3.4.1.2. Catering to the aversion of others

This is the form of discrimination that has, for instance, taken place when women have been fired for wearing Islamic headscarves because customers did not like being attended by someone showing Islamic religious signs.[242] Like the previous type of discrimination, this is a form of taste-based discrimination as it stems from aversion towards the protected group. However, when business owners make this type of discriminatory decision, they do so based on what is more efficient for their business interests. This difference between animus-based discrimination and discrimination that takes place in order to comply with customers' prejudices is highly relevant when applying the anti-discrimination legal framework, seeing as this type of discrimination can be considered to fall within the scope of indirect discrimination, thus leaving the door open to justifications.[243]

3.4.1.3. Cartel model discrimination[244]

Another form of "efficient" discrimination is the one that follows the logic of discriminating against the historically oppressed group in order to provide economic gains for members of the advantaged group. By excluding or relegating members of the protected group to worse positions in the market, the group that already has an economic advantage furthers their benefit.[245]

3.4.1.4. Statistical discrimination

Statistical discrimination is the only form of discrimination that can be justified in the public sector, particularly when it comes to law enforcement regulations and policies.[246] Statistical

---

[242] AGUILERA RULL, A., *Contratación y Diferencia…*, *cit.*, 2013, p. 88; CJEU Judgment 14th March 2017, C-188/15, Asma Bougnaoui and Association de défense des droits de l'homme (ADDH) v. Micropole SA; DONOHUE, J. J., "Antidiscrimination law", in POLINSKY, A. M. & SHAVELL, S., *Handbook on Law and Economics*, vol. 2, Amsterdam, North-Holland (Elsevier), p. 1394.

[243] The CJEU has addressed this issue in two landmark cases that will be studied in Section II. See CJEU Judgments 14th March 2017, C-157/15, Samira Achbita and Centrum voor gelijkheid van kansen en voor racismebestrijding v. G4S Secure Solutions NV and C-188/15, Asma Bougnaoui and Association de défense des droits de l'homme (ADDH) v. Micropole SA.

[244] DONOHUE, J. J., "Antidiscrimination law", *cit.*, 2007, pp. 1409-1410.

[245] AGUILERA RULL, A., *Contratación y Diferencia…*, *cit.*, 2013, pp. 88-89; DONOHUE, J. J., "Antidiscrimination law", *cit.*, 2007, pp. 1409-1410.

[246] While reverse or positive discrimination can also be justified, there is a widespread debate on whether these types of measures should be conceptualised as a form of discrimination, given the negative connotations of the term "discrimination" and the fact that the objective of said measures is to redress the harms caused by structural discrimination. I therefore prefer to analyse what some may label as "positive or reverse discrimination" as part of affirmative or positive action within the substantive equality framework. For a

discrimination in theory takes place when a legitimate aim is pursued. This legitimate aim, which can for example be "promoting the best workers", can however indirectly lead to discriminatory outcomes.

It is commonly agreed that statistical discrimination is more prevalent than animus-based discrimination.[247] Statistical discrimination is therefore not based on conscious prejudices held against certain disadvantaged groups and will therefore be apparently rational as they follow certain stereotypes that, in some cases, do in fact reflect the reality of many members of a group.[248]

This form of discrimination may sometimes turn out to render more efficient results than non-discriminatory options,[249] which means that, in said instances, the trade-off between equality and efficiency must be addressed. However, the problem with addressing discrimination as a trade-off between equality and efficiency is the fact that efficiency as a goal is generally framed in economic terms and consequently offers a very narrow scope of analysis that tends to fall within the structure of dominant narratives.[250] Statistical discrimination will sometimes produce outcomes that do accurately identify certain patterns that confirm prejudices against historically oppressed groups. However, even then, said patterns generally result from the situation of disadvantage suffered by those groups.[251]

Statistical discrimination can fall under the scope of direct discrimination when membership to a protected group is used as a proxy for determining an individual quality or fact, such as

---

detailed explanation on why no type of affirmative action should be categorised as "positive discrimination" see BARRÈRE UNZUETA, M. A., "Igualdad y 'discriminación positiva'…", *cit.*, 2003b, pp. 1-27.

[247] STRAHILEVITZ, L., "Privacy versus antidiscrimination", *cit.*, 2008, p. 373.

[248] STRAHILEVITZ, L., "Privacy versus antidiscrimination", *cit.*, 2008, pp. 365-366: "To illustrate how statistical discrimination plays out in contemporary society, suppose a person charged with hiring a sales clerk wants to avoid employing someone with a criminal background […] assuming the decisionmaker lacks reliable access to information about applicants' criminal records, he might choose to hire a Caucasia female over an equally qualified African-American male, based on the relatively high percentage of African-American males and the relatively low percentage of Caucasia females who are involved in the criminal justice system."

[249] NORMAN, P., "Statistical discrimination and efficiency", *The Review of Economic Studies*, vol. 70, No. 3, 2003, pp. 615-616.

[250] HADFIELD, G. K., "Feminism, fairness, and welfare: an invitation to feminist law and economics", *Annual review of law and social science*, vol. 1, 2005, p. 295: "All of modern corporate law and economics, for example, focuses on the efficiency of corporate governance, understood as maximization of returns to share-holders. Corporate governance, however, has implications for the organization of the workplace, which has implications—profound implications—for the provision of caring labor. The unique attributes of caring labor, however, […] imply that we cannot simply maximize the wealth of shareholders—the profits of firms—and then redistribute income to achieve the socially preferred level of caring labor. Similarly, workplace regulations—hours legislation, parental leave policies, minimum wages, and so on—have to be evaluated not merely in light of efficiency concerns but also in light of ultimate social preferences for the production and quality of caring labor."

[251] MORAN, R., "Whatever happened to racism?", *St John's Law Review*, vol. 75, No. 4, 2005, pp. 899-927.

past criminal convictions. For example, if an employer thinks members of racial minorities are more likely to have been previously convicted she will discriminate against black and Hispanic applicants, not because she dislikes their race-group but because she uses race as a proxy for another data point.[252] This type of statistical discrimination could fall under the category of direct discrimination because individuals are being treated less favourably on the basis of race.

Statistical discrimination can fall within the scope of indirect discrimination when the variables used to determine the target outcome are shaped in such a way that it harms members of the protected group. This is especially relevant when it comes to analysing algorithms, seeing as the variables coded into decision-making processes can easily lead to this form of discrimination. For example, including the number of times an individual has been stopped by the police as a relevant feature in predicting recidivism risk might seem as a neutral criterion but does in fact disadvantage racial minorities.[253]

However, the fact that an apparently neutral and objective criterion is used in order to predict certain outcomes does not necessarily mean that it is the best or most accurate way in which to measure a social phenomenon. It is therefore also important to differentiate between accurate and inaccurate statistical discrimination. Inaccurate statistical discrimination results from using protected group membership as a proxy in a decision-making process. For instance, negative stereotypes regarding black people that stem from the representativeness heuristic[254] can be applied to hiring decisions.[255]

Inaccurate statistical discrimination stems from decisions being made according to a series of social norms that place dominant groups at an advantage while not necessarily accurately depicting reality or predicting outcomes in a given scenario. For example, when granting loans, a bank may provide more weight to individuals' salary than to their saving capacity in order to determine their creditworthiness. Choosing one variable over the other stems from a series of social structures that attribute more value to high-earners than to non-spenders. This will benefit men more than women as a result of the gender wage-gap. However, this does not mean that the variable used is actually indicative of the probability that the individual will

[252] STRAHILEVITZ, L., "Privacy versus antidiscrimination", *cit.*, 2008, pp. 363-381.
[253] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, pp. 24-27.
[254] TVERSKY, A. & KAHNEMAN, D., "Judgment under uncertainty: heuristics and biases", *Science*, vol. 147, No. 4157, 1974, pp. 1124-1127.
[255] BOHREN, J. A. *et al.*, "Inaccurate statistical discrimination", *NBER* Working Paper No. 25935, 2019, p. 3. Available on 3rd July 2019 at: https://www.nber.org/

fully comply with the terms and conditions of the loan. Hence, this form of statistical discrimination would be inaccurate if there is another better predictor for creditworthiness that does not discriminate against women. Otherwise, that is, if there is no better predictor for creditworthiness, we would be facing a case of accurate statistical discrimination.

Accurate statistical discrimination may take place, for example, if a police force decides to increase surveillance on a neighbourhood that is mostly populated by families of Roma origin because there is a widespread believe that said communities tend to engage in illegal activities at higher rates than other ethnic populations. When tested for possible biases this believe may result to be accurate. In these cases, it is important to take into consideration two elements. Firstly, it is very hard to test for all possible forms of bias. Inaccurate discriminatory decisions can result from elements that seem neutral at face-value and the alternatives to the biased shaping of a decision may be almost impossible to detect. Secondly, even if a discriminatory decision is accurate, it is necessary to always work under the presumption that the behaviour of members of protected groups is the result of the construction of society through processes that have historically oppressed and subordinated said groups, placing them at a position of disadvantage and, in many cases, marginalisation.

3.4.2. Equality policies and prohibitions to discriminate are necessary

The previous section shows how discriminatory treatment can sometimes offer benefits in the short-term, meaning that discrimination can, in some cases, be economically reasonable and therefore market powers cannot eliminate discriminatory attitudes on their own. For instance, choosing not to hire young women due to the risk that they might get pregnant may make sense for firms that would have to incur in a series of extra costs if a female employee went on maternity leave, such as training the employee's replacement. However, these discriminatory treatments are generally only beneficial in the short-term. Following from the previous example, it is not socially or economically beneficial to disincentivise parenthood or limit the chances that certain population groups have in accessing the labour market. In these cases, public intervention is necessary in order to guarantee a certain degree of protection to the rights to equality and non-discrimination and to protect other goals that are socially and/or economically beneficial in the short or long run.

The previous statements must not be taken to mean that public intervention can only be justified when equality is more efficient than discriminating against the members of certain

groups. The protection of equality and non-discrimination is necessary as long as there is discrimination. The level of protection is what will vary depending on specific situations and ideological stances.

3.5. THE *STRICTU SENSU* PROPORTIONALITY TEST AND THE NEED TO INTRODUCE STRUCTURAL DISCRIMINATION INTO THE EQUATION

The institutionalisation of oppression throughout history led to the development of policies that openly discriminated against certain groups of people. Clear examples of this can be found in racial segregation laws or regulations limiting women's property rights. However, this kind of direct discrimination with no justification other than the fact that a certain group of people is considered inferior in some way to those in power has become unacceptable in Western societies, resulting in the progressive elimination of such clearly institutionalised forms of discrimination.[256] However, the persistence and pervasiveness of dominant narratives results in the perpetuation of social power structures that continue to discriminate against the members of groups that have been historically oppressed.

The objective of this section is to explain the historical evolution and current pervasiveness of a series of narratives that are present in democratic societies and which systematically place the members of certain groups in subordinated social positions. That is, the following pages focus on generally explaining the way in which structural discrimination is built and works. This reality must be considered when carrying out any proportionality test involving the conflict between equality and other competing values. In this vein, we establish the need for incorporating structural discrimination as a category of analysis in the proportionality test whenever the equality and non-discrimination of disadvantaged groups are being considered, whether it is for the purpose of judging a particular case or determining whether a rule or policy accommodates to legal standards.

3.5.1. Unequal positions of departure

Individuals' ignorance is a constant in social contract theory. RAWLS' hypothetical theory of the social contract drawn from the perspective of fairness focuses on the development of principles of justice and, thus, principles by which society is ruled in a primary state of

[256] MORAN, R., "Whatever happened to racism?", *cit.*, 2005, pp. 899-900.

equality[257] in which the 'veil of ignorance' prevents individuals from being aware of their social position and identity.[258] Highlighting the strength of this idea in RAWLS' work is especially important because he develops his theory of justice from the standpoint of fairness. The 'original position' of ignorance is necessary in order to achieve fair principles of justice.[259] This leads us to a very simple conclusion: since the primary state of equality never occurred, society has been built through norms that privileged the members of groups that have historically held power.

Society has therefore been built, firstly, through differentiating between groups of people and ascribing certain values and characteristics, which conform an identity, to one group as opposed to the other. Secondly, a social structure that hierarchizes between those two (or more) identities has been built, placing one group's values as the nucleus upon which society is built while placing the other group or groups' attributes as secondary. This leads to discussions regarding any issue to be framed from the perspective of the dominant identity's values, which are presented as a neutral framework. As it will be discussed later on, specifying and addressing problems from the standpoint of dominant narratives is an issue that is reproduced in algorithmic decision-making, thereby underpinning the biases and stereotypes that harm traditionally oppressed groups.

3.5.2. The liberal notion of the free autonomous individual

3.5.2.1. Individuals in classic contractualism

A common thread throughout classic social contract theory is the fact that the individual who enters the contract, and thus the prototypical human upon which social norms are built, is defined in terms of freedom.[260] It is the free individual who decides to enter the social contract as well as the rules it contains.[261] Such free individuals are characterised as rational, independent and physically, mentally and economically autonomous[262]. Furthermore, an idea

[257] RAWLS, J., *A Theory of Justice*, Cambridge (Massachusetts), Harvard University Press, 1999 (revised edition), p. 10.
[258] *Idem*, p. 118.
[259] *Idem*, p. 15.
[260] PATEMAN, C., *The Sexual Contract*, Stanford, Stanford University Press, 1988, p. 39.
[261] *Idem*, p. 40.
[262] BERNS, S., "Liberalism and the privatised family: the legacy of Rousseau", *Res Publica*, vol. 11. No. 2, 2005, pp. 127-128; KANT, I., *Grounding for the Metaphysics of Morals*, *cit.*, 1993 (first published in 1785), p. 50, par. 447-448.

that is clearly reflected in LOCKE's *Second Treatise of Government*[263] and that has become an essential and central idea in Western liberal political thought,[264] is the fact that these attributes, as expressions of freedom, are presented as being acquired largely through property,[265] and thus secured by recognising property rights.[266]

Moreover, while ROUSSEAU believes that the role of the state is not just to protect individuals' property but that property does in fact belong to the collective,[267] who must then distribute it, he does consider the recognition of property rights as part of the social contract.[268] In any case, the idea of freedom through property has become an essential and central idea in Western liberal political thought,[269] which, as will be discussed in the following pages, is highly responsible for the narrative underpinning the oppression of certain groups.

A second common element in the first accounts of contractualism is the exclusion of non-whites and women, as well as other groups, from entering the contract.[270] In order to justify these exclusions, contractualists built a narrative through opposition; the excluded groups cannot be part of the contract because they are savage and thus not rational (non-whites),[271] because they are weak and emotional and consequently not rational or independent (women),[272] or because they need more resources than those they have to offer which makes them inadequate to become full co-operators in the contract (individuals with disabilities).[273]

It is important to point out that not all early contractualists drew these ideas from an essentialist perspective. For example, Rousseau defends that women must be educated to

---

[263] LOCKE, J., *Two Treatises of Government. Second essay: Concerning the True Original Extent and End of Civil Government*, pp. 115-126, par. 24-51. Available on 15th April 2019 at: http://www.yorku.ca/
[264] NEDELSKY, J., "Reconceiving autonomy: sources, thoughts and possibilities", *Yale Journal of Law & Feminism*, vol. 1, 1989, pp. 15-16.
[265] *Idem*, p. 22.
[266] MILLS, C., *The Racial Contract*, Ithaca, Cornell University Press, 1997, pp. 31-32.
[267] ROUSSEAU, J. J., *The Social Contract*, 2017 (first published in 1762), p. 10. Available on 15th April 2019 at: https://www.earlymoderntexts.com/
[268] *Idem*, p. 9.
[269] NEDELSKY, J., "Reconceiving autonomy…", *cit*., 1989, pp. 15-16.
[270] CLIFFORD, S., "The capacity contract: Locke, disability, and the political exclusion of "Idiots", *Politics, Groups and Identities*, vol. 2, No. 1, 2014, pp. 90-103; MILLS, C., *The Racial Contract*, *cit.*, 1997, pp. 64 *et seq*.; PATEMAN, C., *The Sexual Contract*, *cit.*, 1988, p. 41.
[271] MILLS, C., *The Racial Contract*, *cit.*, 1997, pp. 64 *et seq*.
[272] RODRÍGUEZ RUIZ, B., "¿Identidad o autonomía? La autonomía relacional como pilar de la ciudadanía democrática", *Identidad, Derecho y Política*, No. 17, 2013, p. 78; WOLLSTONECRAFT, M., *A Vindication of the Rights of Woman*, 2017 (first published in 1974), pp. 5, 24. Available on 15th April 2019 at: https://www.earlymoderntexts.com/
[273] SILVERS, A. & STEIN, M. A., "Disability and the social contract", *The University of Chicago Law Review,* vol. 74, 2007, p. 1616.

become dependent on men, self-abnegated and irrational.[274] Therefore, although he defends the construction of gender roles as necessary in building a good civil society he does recognise it as a social construction.[275]

3.5.2.2. Liberal equality

Western countries now generally recognise formal equality to all their citizens and, in 20th and 21st century liberal theory, individuals who were initially excluded are increasingly accepted as full citizens and thus automatically introduced as part of the contract. This contract provides every individual with a nucleus of legally protected essential negative freedoms.[276] Negative freedom constitutes the lack of interference by the state in individuals' affairs, providing them with the possibility of carrying out (or abstaining from) certain actions.[277] It is the idea that the role of the state is to simply guarantee the preservation of a certain area of privacy within individuals' lives which allows them to freely develop their personhood.[278] It is not that an individual is completely free from all coercion, but that said coercion must be reduced to a minimum that is aimed towards preventing that the exercise of one individual's freedom unjustifiably limits another's.[279]

This notion of freedom is built through a series of civil and political rights, such as religious freedom or freedom of expression, which constitute the core of liberal democracies.[280] These rights provide individuals with a sphere of protection against the interference of political authority and public powers.[281] In addition, theorists pertaining to the different families of political liberalism also reclaimed equality in some form or another.[282] The most basic form of equality reclaimed by political theorists is equality before the law and political equality,

[274] BERNS, S., "Liberalism and the privatised family…", *cit.*, 2005, p. 148.
[275] *Idem*, p. 146: "Rousseau, having acknowledged that male and female are sufficiently similar in innate attributes to be capable of reciprocity, seeks to ensure that this will not be realised."
[276] BERLIN, I., *Liberty*, Oxford, Oxford University Press, 2002.
[277] *Idem*, pp. 169-178.
[278] *Idem*, pp. 178-181.
[279] HAYEK, F., *New Studies in Philosophy, Politics, Economics and the History of Ideas*, London, Routledge, 1978, pp. 132-133.
[280] POYANCO BUGUEÑO, R. A., "Los derechos sociales y la libertad: un análisis problemático", *Derecho Público Iberoamericano*, No. 9, 2016, p. 43.
[281] *Ibidem.*
[282] SEN, A., *The Idea of Justice*, *cit.*, 2009, pp. 291-292.

which are absolutely essential in order to guarantee that the sphere of negative freedoms is granted to all individuals.[283]

However, as egalitarian liberals recognise, the idea of freedom must be linked to certain minimum material conditions and real opportunities since most freedoms can only be exercised if certain pre-existing conditions are given.[284] This relationship is especially relevant because the distribution of material conditions and real opportunities is the result of society being built upon structures that have systematically disadvantaged certain groups.[285] Thus, non-interference in the distribution of resources in society does not guarantee maximum freedom to all but only to those who are born into certain social positions and have certain endowments.

The level to which government intervention in organising this distribution has been argued in political and economic thought has depended on the perspective from which the theory has been addressed. For example, within the liberal spectrum of thought, libertarian approaches, such as NOZICK'S, have heavily criticised most forms of redistribution, equating redistribution through taxation with forced labour.[286] However, many liberal egalitarians, such as RAWLS or DWORKIN[287] have advocated for different forms of compensation built within the framework of distributive justice, defending the idea of a fair distribution of goods, rights and burdens across society that underlies the welfare state.[288] This research mostly focuses on RAWLS' egalitarian approach to the notion of the liberal individual and his theory of distributive justice due to the fact that this stream of thought has gained general acceptance and occupied a central spot in the debate on inequality, freedom, efficiency and their role in structuring societies throughout the late 20th and early 21st Centuries.[289]

RAWLS' version of distributive justice is materialised in two principles which, he argues, would be chosen by rational individuals in a state of absolute ignorance regarding their

[283] ATIENZA, M., *El Sentido del Derecho*, Barcelona, Ariel, 7ª ed., 2018, p. 192; SEN, A., *The Idea of Justice*, *cit.*, 2009, pp. 291-292.
[284] ATIENZA, M., *El Sentido del Derecho*, *cit.*, 2018, p. 192; POYANCO BUGUEÑO, R. A., "Los derechos sociales y la libertad…", *cit.*, 2016, p. 42; RAWLS, J., *A Theory of Justice*, *cit.*, 1999, p. 87;
[285] CROCKER, L., *Positive Liberty: An Essay in Normative Political Philosophy*, The Hague, Martinus Nijhoff Publishers, 1980, p. 9.
[286] NOZICK, R., *Anarchy, State and Utopia*, Oxford, Blackwell, 1974, p. 169. NOZICK does, however, admit some form of redistribution, albeit minimal and circumstantial, through the principle of rectification (p. 231).
[287] RAWLS, J., *A Theory of Justice*, *cit.*, 1999, *cit.*; DWORKIN, R., *Sovereign Virtue*, Cambridge, MA, Harvard University Press, 4th ed., 2002.
[288] ROSENFELD, M., "Affirmative action, justice, and equalities: a philosophical and constitutional appraisal", *Ohio state Law Journal*, vol. 46, No. 4, 1985, p. 859.
[289] ATIENZA, M., *El Sentido del Derecho*, *cit.*, 2018, p. 211.

identity and social position.[290] The first is the basic principle upon which all liberal interpretations of justice rest, which is the establishment of a system of freedoms that equally applies to all individuals.[291] The second principle, is divided in two parts. The first part, labelled "the difference principle",[292] contends that individuals should be equally assigned the most basic rights and burdens and inequalities should be shaped so that they benefit the least advantaged in the long term.[293] The second part is "formal equality of opportunity in that all have at least the same legal rights of access to all advantaged social positions".[294] RAWLS gave priority to the freedoms contained in the first principle over equality of opportunity and the difference principle in the sense that the former could not be forfeited in favour of the latter.[295]

RAWLS' defence of forms of inequality that would generate an expectation of future better outcomes for those who are worse off brings about the discussion of the extent to which the difference principle is related to trickle-down economics. For RAWLS, the main principle grounding justice is equality. Inequalities that benefit better-off individuals, while accepted, are only to exist if they result in better outcomes for the less well-off individuals and groups.[296] Conversely, trickle-down economics, also known as supply-side economics, focuses on producing a general improvement of the economy as well as a decrease in deficit.[297] Supply-side economics posits inequality as a given element of the economic organisation of society that is necessary in order to incentivise behaviours that will lead to general economic growth. They argue that not forcing redistribution will create incentives for the rich to generate more wealth which will trickle-down to middle classes and, subsequently,

[290] RAWLS, J., *A Theory of Justice*, *cit.*, 1999, p. 17.
[291] *Idem*, p. 53.
[292] *Idem*, p. 65 *et seq*.
[293] *Idem*, p. 39: "…economic and social inequalities are to be judged in terms of the long-run expectations of the least advantaged social group"; p. 68: "The inequality in expectation is permissible only if lowering it would make the working class even more worse off".
[294] RAWLS, J., *A Theory of Justice*, *cit.*, 1999, p. 62.
[295] *Idem*, pp. 24-25.
[296] NATHANSON, S., Economic Justice, New Jersey, Prentice Hall, 1988. Available on 24th April 2019 at: http://www.woldww.net/
[297] KRAYNAYA, D., "Economics and jurisprudence: is John Rawls' difference principle just another form of supply side economics and can it be applied effectively in modern society?", *Manchester Student Law Review*, vol. 1, 2012, p. 54.

the poor.[298] However, redistribution and equality is, by no means, the main objective they aim to achieve.[299]

Furthermore, RAWLS' intention is to set up a theory that establishes the basic principles upon which society can work in order to shift power structures to also encompass the particular interests and needs of traditionally oppressed groups as he very clearly reflects by stating:

> " [W]e may reject the contention that the ordering of institutions is always defective because the distribution of natural talents and the contingencies of social circumstance are unjust, and this injustice must inevitably carry over to human arrangements. Occasionally this reflection is offered as an excuse for ignoring injustice, as if the refusal to acquiesce in injustice is on a par with being unable to accept death. The natural distribution is neither just nor unjust; nor is it unjust that persons are born into society at some particular position. These are simply natural facts. What is just and unjust is the way that institutions deal with these facts."[300]

Nonetheless, since the main difference that can be found between RAWLS' theory and supply-side economics is more theoretical than practical, defenders of trickle-down economics have been able to present their theory as subsumed within RAWLS' difference principle. Proponents of supply-side economics argue that allowing the wealthy to accumulate more resources will eventually lead to a general welfare improvement for all members of society. It is thus not easy to argue that these theories do not fall under the mandates of the difference principle.[301]

Establishing this relationship between the main theoretical proposal of liberal egalitarianism and non-interventionist economic theories is relevant in order to convey the extent to which RAWLS' theory can be approached from different political and economic perspectives and that it can be easily adopted by more conservative forms of liberalism.[302] Moreover, and what is particularly relevant towards the theoretical framework here presented, even when the difference principle is used as the base for more egalitarian theoretical or practical proposals for a fair and just society, these are still constructed in a way that presents institutions as neutral entities and denies or ignores the particularities and specific forms of oppression

[298] REIFF, M. M., "The difference principle, rising inequality, and supply-side economics: how Rawls got hijacked by the right", *Revue de philosophie économique*, vol 13, No. 2, 2012, p. 127.
[299] KRAYNAYA, D., "Economics and jurisprudence…", *cit.*, 2012, p. 54.
[300] RAWLS, J., *A Theory of Justice*, *cit.*, 1999, p. 87.
[301] REIFF, M. M., "The difference principle, rising inequality, and supply-side economics…", *cit.*, pp. 128-130.
[302] *Idem*, pp. 129.

undergone by certain groups.[303] Said proposals do not deny the existence of unequal positions of departure but generally consider that by extending equal rights to all and providing some minimum mechanisms to correct for the disadvantaged position of some individuals are sufficient to guarantee a necessary degree of freedom to all. More importantly, they generally accept the existence of inequalities, providing more importance to values that compete with equality in certain contexts.

3.5.2.3. Liberal thought and the perpetuation of group disadvantage

The perpetuation of structures of oppression in contemporary liberal thought and democracies thus results from the inclusion in the social contract of those who were traditionally excluded. While this may seem counterintuitive, it is the consequence of predicating a universality of rights built on the same notion of the individual put forward by classic contractualist theory.[304] This construction is especially perverse because, while current 'social contracts' apparently comprise all people and their different identities, the framework has not shifted in order to include the particularities and values of those groups that make them different from the individuals that have always been included in the social contract:[305] the autonomous, cis, heterosexual, white male of a certain socioeconomic status.[306] In this sense, RAWLS' theory of distributive justice, while placing equality as one of its main elements, is still constructed from the perspective of the dominant narrative that it aims to modify.[307]

One of the main reasons why liberal thought generally fails to recognise how the narratives of domination that have been present in our societies through history are now embedded in what apparently seem neutral institutions and norms is the fact that the unit of analysis is the individual, who is viewed and understood as a free person. Since the free individual is not

[303] BARRANCO, M. C., "La concepción republicana de los derechos en un mundo multicultural", in DEL REAL ALCALÁ *et al.* (coords.), *Derechos Fundamentales, Valores y Multiculturalismo*, Madrid, Dykinson, 2005, p. 17; OKIN, S. M., "Gender, justice and gender: an unfinished debate", *Fordham Law Review*, vol. 72, No. 5, 2004, pp. 1537-1567.

[304] BARRÈRE UNZUETA, M. A., "Iusfeminismo y derecho antidiscriminatorio…", *cit.*, 2008, pp. 55-56; PATEMAN, C., *The Sexual Contract*, *cit.*, 1988, p. 42: "…almost all the classic writers held that natural capacities and attributes were sexually differentiated. Contemporary contract theorists implicitly follow their example, but this goes unnoticed because they subsume feminine beings under the apparently universal, sexually neuter category of the 'individual'"; PATEMAN, C. & MILL, C., *Contract & Domination*, Cambridge, Polity Press, 2007.

[305] BERNS, S., "Liberalism and the privatised family…", *cit*., 2005, pp. 154-155; PATEMAN, C., *The Sexual Contract*, *cit.*, 1988, p. 42

[306] BARRANCO, M. C., *Diversidad de Situaciones y Universalidad de los Derechos*, Dykinson, Madrid, 2011, pp. 14-15; BARRÈRE, M. A., "Igualdad y 'discriminación positiva'…", *cit.*, 2003b, p. 7; SALOMÉ, L. M., "La discriminación y algunos de sus calificativos…", *cit.*, 2017, p. 282.

[307] MILLS, C., *The Racial Contract*, *cit.*, 1997, p. 77.

always analysed in the context and in relation to the society she lives in and the identity-groups she might pertain to, many liberal accounts of social phenomena, structures and institutions fall short as they do not manage to draw a comprehensive analysis of the external constraints to which individuals are sometimes subjected to when exercising freedom.

3.5.3. Dominant narratives of oppression: identifying disadvantage

The fact that current anti-discrimination legal frameworks almost solely recognise instances of specific direct or indirect discrimination and practically do not consider the existence of structural discrimination is perhaps the paramount demonstration of the way in which liberal thought has pervaded all social institutions, including those aimed towards restructuring existing imbalances in power. This particular shortcoming in anti-discrimination law, which will be further analysed later on, conveys the insufficiencies of failing to analyse individuals in the social context in which they develop.[308]

Throughout the following pages several narratives of oppression will be discussed. While no specific cases of discrimination will be addressed, the objective is to show the way in which domination discourses create an environment which favours the general social disadvantage of members of certain groups as well as the appearance of specific situations of discrimination.

The categories chosen are mainly those that are generally protected by non-discrimination clauses in democratic declarations of rights and Constitutions. However, it is important to keep in mind that said clauses are mostly open-ended and allow courts and public bodies to identify especially disadvantaged groups that are not yet included.[309]

Categorising a group as traditionally oppressed or disadvantaged and explaining the way in which said oppression has been constructed and is still present in society is vital for two reasons. On the one hand, it is necessary to establish which groups have been traditionally discriminated against and how said discrimination and disadvantage has been constructed and is currently articulated in order to understand the way in which algorithmic discrimination operates, as it follows very similar logics to traditional forms of discrimination. On the other

[308] BARRÈRE UNZUETA, M. A. & MORONDO TARAMUNDI, D., "Subordiscriminación y discriminación interseccional: elementos para una teoría del derecho antidiscriminatorio", *Revista de Filosofía Jurídica y Política*, vol. 45, 2011, pp. 15-42.
[309] AÑÓN ROIG, M. J., "Principio antidiscriminatorio y determinación de la desventaja", *cit.*, 2013b, p. 135.

hand, it is essential to determine which groups have been traditionally oppressed, to explain their history of discrimination and the specific situations of past and present disadvantage, as it allows them the highest level of protection offered by the equality and non-discrimination clauses present in constitutions and international legal instruments aimed towards the protection of human rights.[310]

When explaining the way in which the different narratives of domination have been built, disadvantaged groups are treated as unitary elements. This does in no way mean that all members of the group are considered to be identical beings and that their individuality is denied, but that the structural discrimination that is built through narratives of domination has an undeniable collective element. While acknowledging that the social structures of discrimination are mostly materialised in the way they specifically impact individuals, the following pages aim to highlight that said specific impact results from individuals' pertaining to disadvantaged groups.[311]

In addition, while the main focus here is set on the disadvantaged groups that result from the construction of narratives of domination, it is also important to keep in mind that the harms caused by discrimination and the groups that suffer them are not static or unidirectional. Dominant narratives can also harm members of the group in power. For example, gender roles do not simply oppress women but also establish very clear determinations of what a "man" should be, punishing men who in any way deviate from the norm. The pervasive nature of narratives of oppression also means that they are constantly reinforced and perpetuated by society as a whole and it is therefore generally not possible to clearly establish a dichotomy between oppressed and oppressors, especially considering that practically all individuals are either born into a disadvantaged group or enter into it at some point during their lives.

The main categories studied in the following pages are: the subordination of women, non-whites and lower socioeconomic strata. The reason why these are the main categories analysed is that most of the examples of algorithmic discrimination that will be studied in this dissertation are instances of racism, sexism and classism. It is important to highlight that this dissertation aims to draw an intersectional critical framework and does in no way aim to be

[310] *Idem*, p. 131-132.
[311] *Idem*, p. 134.

built from the perspective of "oppression Olympics",[312] which results in the hierarchisation of the different forms of disadvantage and the idea that a certain form of discrimination is greater that another and more public resources should be therefore focused on it. Nonetheless, as it was indicated when addressing intersectional discrimination, purely for research purposes, it is necessary to focus on more general discriminated groups rather than trying to analyse all possible intersecting identities.

3.5.3.1. Gender roles and male domination

i) *Introduction to gender-based discrimination*

Throughout history, the female identity has been constructed as secondary with respect to the male identity, excluding women from participation in any form of power.[313] The creation of this male-dominant structured society has been the product of dividing society in two differentiated spheres; public and private, having men occupied the former while women were relegated to the latter.[314]

Men and women are assigned differentiated values, which are, in turn, associated to either the public or private sphere[315] and which result in the sexual division of labour.[316] Thus, while the female identity is built upon values related to care, solidarity, cooperation, empathy, dependence and irrationality, among others,[317] the construction of the male identity rests on values such as aggressiveness, competitiveness, rationality, objectivity and independence.[318]

In the construction of "the sexual contract",[319] these two spheres are hierarchized so that female values are subordinated to male values, thereby justifying the exercise of power of men over women.[320] Amongst the many ways in which men have exercised power over women through this female-male or masculinity-femininity dichotomy is the fact that

[312] MAHALINGAM, R., BALAN, S. & HARITATOS, J., "Engendering immigrant psychology: an intersectionality perspective", *Sex Roles*, vol. 59, 2008, p. 326.
[313] LERNER, G., *The Creation of Patriarchy*, Oxford, Oxford University Press, 1987, p. 5.
[314] PATEMAN, C., *The Sexual Contract*, *cit.* 1988, p. 11.
[315] RODRÍGUEZ RUIZ, B. & RUBIO MARÍN, R., "Constitutional justification of parity democracy", *Alabama Law Review*, vol. 60, No. 5, 2009, pp. 1181.
[316] LERNER, G., *The Creation of Patriarchy*, *cit.*, 1987, p. 24.
[317] FERNÁNDEZ RUÍZ-GÁLVEZ, E., *Igualdad y derechos humanos*, Madrid, Tecnos, 2003, pp. 160-161.
[318] GALLEGOS ARGÜELLO, M. C., "La identidad de género: masculino versus femenino", in SUÁREZ VILLEGAS, J. C., LIBERIA VAYÁ, I. & ZURBANO-BERENGUER, B. (coords.), *I Congreso Internacional de Comunicación y Género. Libro de Actas: 5, 6 y 7 de marzo de 2012. Facultad de Comunicación. Universidad de Sevilla*, Sevilla, Universidad de Sevilla, Facultad de Comunicación, 2012, p. 713.
[319] PATEMAN, C., *The Sexual Contract*, *cit*, 1988.
[320] *Idem*, pp. 50-51.

women's bodies, especially their sexuality and reproductive capacity, have been portrayed as, and actually become, a marketable commodity.[321]

While women managed to enter the public sphere and, to a certain extent, occupy positions of power in society decades ago in most Western countries, this development did not in turn lead to the entry of men in the private sphere and, in fact, most household-related work is still carried out by women.[322] Additionally, the incorporation of women to the public sphere has been, in many cases, characterised by a division of labour along female and male occupations, being the latter perceived as superior.[323]

ii) *Perpetuating gender roles through the political theories that shape our society*

Society is still heavily constructed upon male values and policymaking, power structures and the provision of goods and services are generally designed from an androcentric perspective.[324] This is clearly reflected in some of the main political thought that has shaped Western democracies throughout the 20th and 21st Centuries. For example, in RAWLS' *Theory of justice*,[325] women are included as equal subjects in the social contract. However, by constructing a universal notion of the 'individual' fundamentally based on androcentric values, he failed to address the particularities of the oppression suffered by women as a group as well as their specific values and needs.[326] RAWLS' social contract interprets individuals in the original position as rational beings.[327] Since rationality is a quality that has been associated to masculinity, this construction of the individual that enters the social contract indirectly excludes women who are still associated with the expression of emotion rather than rational thought.[328]

---

[321] LERNER, G., *The Creation of Patriarchy*, *cit.*, 1987; NUSSBAUM, M. C., "Objectification", *Philosophy and Public Affairs*, vol. 24, No. 4, 1995, pp. 249-291; GONZÁLEZ RAMOS, A. M. & TORRADO MARTÍN-PALOMINO, E., "Objectification and marketisation of women: technologies as instrument of violence", *Sociología y Tecnociencia*, vol. 9. No 1, 2019, p.1.
[322] OECD, "Entrenched social norms prevent the equal distribution of caring responsibilities between men and women", March 2018.
[323] BRYANT, G. *The Working Woman Report: Succeeding in Business in the 80's,* New York, Simon & Schuster, 1984, p. 47: "…there is women's work, and there is men's work—and men's is better."
[324] ALBERTÍN CARBÓ, P., CUBELLS, J. & PEÑARANDA, M. C., "A feminist law meets an androcentric criminal justice system: gender-based violence in Spain", *Feminist Criminology*, 2018, p. 4; MUXÍ MARTÍNEZ, Z. M. *et al.*, "¿Qué aporta la perspectiva de género al urbanismo?, *Feminismo/s*, No. 17, 2011, pp. 105-129.
[325] RAWLS, J., *A Theory of Justice*, *cit.*, 1999.
[326] PATEMAN, C., *The Sexual Contract*, *cit.*, 1988, p. 42.
[327] ATIENZA, M., *El Sentido del Derecho*, *cit.*, 2018, pp. 188-190.
[328] OKIN, S. M., "Gender, justice and gender…", *cit.*, 2004, p. 1545.

It could be argued that the theory of justice based on the original position is purely hypothetical and thus it is not necessary to specifically address historical accounts of oppression.[329] However, moving away from critiques based on the methodology of abstraction used by RAWLS,[330] the identification of individuals in the original position with "heads of families",[331] confirms the dominant narrative underlying RAWLS' work.[332] While heads of families are not necessarily men and that there are increasingly diverse families, it is undeniable that the role of the father within the family institution is generally identified with that of the head of the family. Furthermore, considering it is only heads of family who enter the contract in the original position in representation of the rest of their family, the classical division between public and private sphere and thus female subordination is reinforced.[333] In fact, this precise element supports the argument that RAWLS' abstracted individual is the prototypical "liberal man", for the head of the family (the man) who enters the social contract is the rational being, thereby opposing the rationality of individuals who occupy the public sphere with the more emotional (irrational) private sphere of which women take care.

iii) *The essentialising nature of sex and the need for non-assimilation*

The demand that the particularities of women be considered and observed in social design is confronted with the denial of gender roles as 'natural'. The gender roles assigned to each biological sex are social constructs and the essentialising nature of biological sex must consequently be denied. In this context, in which the essentialising nature of biological sex still pervades worldviews shared in society, women are discriminated against as a result of negative stereotyping. For instance, when women are not selected for leadership positions because they are thought to be 'too emotional'.

However, gender roles, while socially constructed, are still very present and real; women do generally exercise caring roles and express more empathy than men. Therefore, feminism sees itself in the obligation of having to deny the essentialising nature of biological sex while having to defend the existence of (socially constructed) differentiated elements that

[329] FOSTER, S. L., "Rawls, race and reason", *Fordham Law Review*, vol. 72, No. 5, 2004, p. 1718.
[330] MATSUDA, M. J., "Liberal jurisprudence and abstracted visions of human nature: a feminist critique of Rawls' theory of justice", *New Mexico Law Review*, vol. 16, No. 3, 1986, p. 613.
[331] RAWLS, J., *A Theory of Justice*, *cit.*, 1999, p. 111.
[332] OKIN, S. M., *Justice, Gender and the Family*, New York, Basic Books, 1997, pp. 94 *et seq*.
[333] BERNS, S., "Liberalism and the privatised family…", *cit*., 2005, pp. 153-154; OKIN, S. M., *Justice, Gender and the Family*, *cit.*, 1997.

characterise women as a group and that lead to their discrimination.[334] Additionally, it is also essential to establish and defend that female-associated values are not negative or worse than male-associated values.

Defending that women are, in fact, different from men is sometimes viewed as paradoxical. However, the existence of this apparent paradox within the feminist scholarship is absolutely necessary in order to convey the demands for a shift in the framework shaping all relationships and structures in both the public and private spheres. One thing is rejecting that gender roles are natural and another thing is neglecting social reality.[335] The lack of observation of women's specific situations and of the specific characteristics of the oppression they suffer leads to reinforcing their social, political and economic discrimination.

Thus, women do generally exercise caring roles at a higher rate than men, but this is mainly the result of the social construction that assigns different roles to men and women and prioritises the former over the latter. However, caring for others is a positive and essential social value. It is therefore necessary to be aware that, in order for women to achieve equality, they must not be treated like men or expected to behave like them. Typically female values must be incorporated into the public sphere and power structures in a position of equality with male values.

This does not mean that there is no biological element in the differences between men and women but that the structural disadvantage suffered by women has been and is the result of a society constructed through their historical subordination through stereotyping and stigmatisation.

iv) *The persistence and pervasiveness of gender-based discriminatory structures*

The lack of real inclusion of women in the public sphere by the philosophers and theorists that have shaped 20th and 21st Century political thought is clearly reflected in the persistence of situations of gender inequality and discrimination. There is still a widespread representation in media outlets and society in general of women as products for male

[334] STEPAN, N. L., "Race, gender, science and citizenship", *Gender & History*, vol. 10, No. 1, 1998, p. 43; NOBLE, S. U., *Algorithms of Oppression…*, *cit*., 2018, p. 95: "At the same time that women reject biological classifications as essentializing features of sex discrimination, they are simultaneously force to organize for political and economic resources and progress on the basis of gender."
[335] LUDVIG, A., "Differences between women?...", *cit.*, 2006, pp. 247-248.

consumption, which leads to gender-based, and specially sexual violence;[336] women are still in charge of most caregiving (unpaid) activities;[337] they undergo harsher economic circumstances and their attempts to improve said situations are hampered, amongst other things, by the discrimination they suffer in the public sphere and by the burden of unpaid labour[338] and, while more women occupy political bodies and other power structures, they are still largely underrepresented,[339] and when they do occupy positions of power they tend to imitate masculine values as these are identified with power roles[340] thereby further deepening the neglect of female identities and particularities within the public sphere.

A particularly pervasive way in which the subordination of women and female values takes place is through the existence of gender-segregation in the workforce which is, in part, to blame for the gender wage gap. There are certain types of employment, such as cleaning and nursing, that involve tasks which are similar to those that have been traditionally exercised by women within the household. These professions are thus mainly occupied by women and are generally undervalued in comparison to other professions that are not associated to female values, thereby resulting in a general difference in the average remuneration of women and men.[341] This dynamic is further reinforced through the fact that women tend to take up part-time jobs at a higher rate than men in order to focus on household activities, which means that the global wealth accumulated by women is far less than that of men.[342]

In addition, by refusing to acknowledge the specific historical oppression exerted upon women in their analysis of social structures and institutions, liberal thinkers fail to recognise the existence (and persistence) of certain structures of oppression. Take, for example, the case of sexual relations and consent. The emergence of women's movements that have

[336] WRIGHT, P. J. & TOKUNAGA, R. S., "Men's objectifying media consumption, objectification of women, and attitudes supportive of violence against women", *Archives of Sexual Behavior*, vol. 45, No. 4, 2016, p. 962.
[337] OECD, "Entrenched social norms prevent the equal distribution of caring responsibilities between men and women", March 2018.
[338] See, in general, RAZAVI, S, "Turning Promises Into Action: Gender Equality in the 2030 Agenda for Sustainable Development", New York, UN Women, 2018. Available on 26th April 2018 at: http://www.unwomen.org/
[339] ATSKE, S., GEIGER, A. & SCHELLER, A., "The share of women in legislatures around the world is growing, but they are still underrepresented", *Pew Research Center*, 18th March 2019.
[340] GONZÁLEZ, F. J., *El Fin del Mito Masculino: La Entrada en el Siglo de la Mujer*, Barcelona, Erasmus Ediciones, 2007, p. 136.
[341] ALKSNIS, C., DESMARAIS, S. & CURTIS, J., "Workforce segregation and the gender wage gap: is 'women's' work valued as highly as 'men's'?", *Journal of Applied Social Psychology*, vol. 38, No. 6, 2008, pp. 1416-1441.
[342] MATTEAZZI, E., PAILHÉ, A. & SOLAZ, A., "Does part-time employment widen the gender wage gap? Evidence from twelve European countries", *Society for the Study of Economic Inequality*, Working Paper 2013-2913, 2013, pp. 28-29.

publicly denounced the extent to which sexual harassment and assault are normalised[343] has been met with the rise of a debate on the significance of sex and consent.

On the one hand, some interpretations of the way in which sexuality is developed and expressed refuse to provide sexual relations with an underlying social significance and are used to reduce the spectrum of what can be considered as non-consensual sexual relationships and even, in some cases, to decrease the severity with which said actions are viewed.[344] On the other hand, even when the deep and intimate meaning that sexual relations have both, in society and for individuals, is recognised, said meaning is not analysed within the power structure of female oppression that has been historically built through male sexual domination.[345] Thus, when analysing consent it is not uncommon for liberal thinkers to not acknowledge the existence of an environment in which women feel intimidated and are therefore unable to respond to certain actions and advances carried out by men. Hence, any attempt to shift the meaning of what actual freedom and consent constitute is criticised by certain liberal accounts that consider the feminist accounts of consent to be patronising and limiting individuals' freedom.

The discussion regarding sexuality and consent, while illustrative of the way in which male power structures have been built and are still present in society, is however not as relevant to the role algorithms play in the perpetuation of female oppression as other elements, such as the subordination of female values. Nonetheless, it is important to highlight how portraying situations that are heavily underpinned by the historical oppression suffered by women from the perspective of the atomistic and free individual that some liberals adopt lacks the more comprehensive understanding of the way in which freedom operates that critical theories, such as feminism, provide.

3.5.3.2. White domination

The narrative built upon the idea of white supremacy in western societies is rests on the creation of the dichotomy of white and non-white, having the former held positions of power,

[343] The #MeToo movement was a stepping stone which helped women to step forward and bring about sexual harassment and assault claims against different men both in and outside the entertainment industry. In addition, for example in Spain, several cases of group-rape have also led many women to provide accounts of their own sexual harassment and assault experiences.
[344] DE LORA, P., *Lo Sexual es Político (y Jurídico)*, Madrid, Alianza Editorial, 2019, pp. 39-40.
[345] JEFFREYS, S., "Kate Millett's sexual politics: 40 years on", *Women's Studies International Forum*, vol. 34, 2011, pp. 76-77.

while the latter were relegated to a secondary position within society. This racial dichotomy traditionally fell along the binary distribution of full persons (white) or non-full or *subpersons* (non-white).[346]

Racial (white) domination is closely related to and in many instances built alongside narratives of ethnic oppression. The European Court of Human Rights expresses the way in which these two narratives interact and can be distinguished in the following terms:

> "Ethnicity and race are related and overlapping concepts. Whereas the notion of race is rooted in the idea of biological classification of human beings into subspecies according to morphological features such as skin colour or facial characteristics, ethnicity has its origin in the idea of societal groups marked by common nationality, tribal affiliation, religious faith, shared language, or cultural and traditional origins and backgrounds."[347]

These two elements work together in many cases given that it is not strange for non-nationals and religious or ethnic minorities to have a different physical appearance to countries' nationals. However, there are also many instances in which narratives of domination driven through ethnicity do not incorporate said physical difference as it is for example the case with language discrimination. This section mostly focuses on those instances in which race and ethnicity act together to generate structures of oppression and specific instances of discrimination.[348]

A common line upon which both male and white domination have been developed is property. The commodification of non-whites has been carried out through centuries of processes of colonisation and enslavement, which were also, largely, what brought by what is known as "the European miracle";[349] the sudden raise at the end of the Middle Ages of what had been a peripheral area into a phase of world dominance, which lasted until the 20th

[346] MILLS, C., *The Racial Contract*, *cit.*, 1997, p. 11.
[347] ECHR Judgment 13th December 2005 (final decision March 13th 2006), 55762/00 and 55974/00, Timishev v. Russia.
[348] Other forms of discrimination related to ethnicity will be briefly addressed later on in this chapter but are not the main focus of the thesis.
[349] MILLS, C., *The Racial Contract*, 1997, *cit.*, p. 33; WILLIAMS, E., *Capitalism and Slavery*, Chapel Hill, the University of North Carolina Press, 1944, p. 52.

century[350] and the consequences of which, developing countries still suffer largely due to neocolonisation.[351]

Another common element with the discourse of female oppression is that white domination has worked by naturalising the differences between whites and non-whites. This domination was partly justified and rationalised through the narrative of white superiority.[352] Non-white identities are therefore built upon values or elements that characterise them as less developed humans than white individuals. For example, in the case of black people, they have been stereotyped as unintelligent, lazy and dishonest.[353]

Furthermore, considering it is essential for the dominant white narrative to present individuals categorised within other races as subpersons and to emphasise the differences between them, the physical aspect of non-whites plays and has played a very important role in developing the discourse of domination, for example, by presenting darker skin colours and certain physical attributes associated to them as ugly.[354]

However, unlike in the case of female domination, racial stereotypes do not lead to a domination built upon personal relationships and the public/private sphere dichotomy.[355] In the case of race/ethnic stereotypes, domination is carried out and analysed as part of the structure of the public sphere.[356] In addition, it is important to keep in mind the existence of specific forms of domination that affect Non-white women, which must be addressed from intersectional perspectives.[357]

The hierarchical race structure thus builds on the idea that non-white people are intellectually inferior and consequently should occupy a secondary or subordinated role with respect to white individuals in society.[358] This leads to the racial organisation of society being largely

[350] JONES, E. L., *The European Miracle: Environments, Economies and Geopolitics in the History of Europe and Asia*, Cambridge, Cambridge University Press, 1981.
[351] MILLS, C., *The Racial Contract*, *cit.*, 1997, p. 36.
[352] *Idem*, p. 33.
[353] NOBLE, S. U., *Algorithms of Oppression…*, *cit*., 2018.
[354] *Idem*, p. 81.
[355] VERLOO, M., "Multiple inequalities, intersectionality and the European Union", *European Journal of Women's Studies*, vol. 13, No. 3, 2006, p. 218.
[356] *Ibidem*.
[357] AÑÓN ROIG, M. J., "Principio antidiscriminatorio y determinación de la desventaja", *cit.*, 2013b, p. 151-152.
[358] MILLS, C., *The Racial Contract*, *cit.*, 1997, p. 59.

derived from the way in which citizenship is conceived as a division along the lines of who belongs to the group and who is the outsider.[359]

These constructions lead to the very real categorisation and hierarchical organisation of humans who develop aligned interests and see many of their needs ignored by society even when they are officially recognised full citizenship because, in the collective imaginary, non-whites are not citizens of that particular country.[360] As it happens with regard to female subordination, the fact that societies are racialised and that racial assignation is a form of oppression[361] requires defending race and ethnicity[362] as social constructions while making it necessary for different ethic or racial groups to claim recognition of their specific situations and particularities.[363] In addition, as with other subordiscriminated groups, the categorisation of society into races or ethnic groups and the negative stereotyping of non-whites, alongside their marginalisation, has been found to generate self-fulfilled prophecies in which certain members of said groups do in fact behave in accordance to the stereotype.[364]

Thus, conceiving human beings from the perspective of the neutral individual as a white person, leads to a political, economic and social structure that discriminates against other races or ethnicities by excluding or discriminating them in the access to resources,[365] not addressing their specific needs and reinforcing stereotypes which associate negative connotations to elements attributed to non-white identities.[366]

Finally, it is important to consider the close relationship between racism and xenophobia, seeing as in many Western countries they go hand in hand and are constructed from very

[359] VERLOO, M., "Multiple inequalities, intersectionality and the European Union", *cit.*, 2006, p. 218.
[360] MORRISON, T., "Making America white again", *The New Yorker*, 14th November 2016. Available on 13th April 2019 at: https://www.newyorker.com/
[361] BASHI, V., "Racial categories matter because racial hierarchies matter: a commentary", *Ethnic and Racial Studies*, vol. 21, No. 5, 1998, p. 966: "Racial identities are obtained not because one is unaware of the choice of ethnic labels with which to call oneself, but because one is not allowed to be without a race in a racialized society. Race is a sociocultural hierarchy, and racial categories are social spaces, or positions, that are carved out of that racial hierarchy."
[362] VERLOO, M., "Multiple inequalities, intersectionality and the European Union", *cit.*, 2006, p. 218: "The label 'race' seems to be constructed as more closely linked to nature, to biology, to being born as belonging to a certain category, while the label 'ethnicity' is constructed as linked more closely to nurture, to culture and geographical roots, but both labels overlap."
[363] STEPAN, N. L., "Race, gender, science and citizenship", *cit.*, 1998, p. 38; NOBLE, S. U., *Algorithms of Oppression…*, *cit*., 2018, p. 84, 95.
[364] APPEL, M., WEBBER, S. & KRONBERGER, N., "The influence of stereotype threat on immigrants: review and meta-analysis", *Frontiers in Psychology*, 2015, pp. 1-15.
[365] SULLIVAN, L. *et al.*, "The racial wealth gap: why policy matters", IASP/Demos, 2015.
[366] WHEELER, S. C. *et al.*, "Think unto others: the self-destructive impact of negative racial stereotypes", *Journal of Experimental Social Psychology*, vol. 37, No. 2, 2001, pp. 173-180.

similar perspectives. In some European countries, such as Spain, many immigrants are still first generation and it is therefore very hard to differentiate between negative attitudes towards these individuals due to their race or national origin.

3.5.3.3. Exclusion based on property: classism and "aporophobia"[367]

Classism or oppression based on socioeconomic status is the narrative which leads to the discrimination of individuals based on their social position. As it was stated when analysing the liberal notion of the individual, in general, property plays a very important part in building the autonomous individual. It is through ownership that he can be economically autonomous and, thus, free.[368] Justifications for suffrage limitations applicable to men were mainly based on this idea.[369] Since property is one of the main elements upon which Western societies are currently constructed, narratives of oppression based on socioeconomic status and the inequality suffered by members of less advantaged social strata is constantly being reinforced.

The dominant narrative of oppression towards less well-off individuals and classes in Western societies has been shaped mainly in two ways. On the one hand, charity has been a driving force in establishing relationships between individuals and groups that belong to different social strata, however, as FISHER LAVELL indicates, "there is a difference between a charitable yet distant view of poor people as unfortunate 'others', and true engagement with social class as a site of oppression."[370] Charity generally perpetuates the image of the poor as "others". Their "otherness" prevents wealthy individuals from seeing them as human beings who deserve full fundamental and social rights and therefore from recognising the existence of a social structure that systematically harms the poor.

On the other hand, wealthier groups and individuals relate to lower social classes through feelings of disgust and condescendence, attributing full responsibility to individuals for their

[367] The term "aporophobia", coined by Adela CORTINA (*Aporofobia, el rechazo al pobre*, Barcelona, Paidós, 2017), refers to a form of classism adapted to the current social stratification that can no longer be separated into clear social classes.
[368] NEDELSKY, J., "Reconceiving autonomy…", *cit*., 1989, pp. 15-16.
[369] PRZEWORSKI, A., "Conquered or granted? A history of suffrage extensions", *British Journal of Political Science*, vol. 39, No. 2, 2009, p. 295.
[370] FISHER LAVELL, E., "Beyond charity: social class and classism in counselling", *Canadian Journal of Counselling and Psychotherapy*, vol. 48, No. 3, 2014, p. 234.

economic situation.[371] In fact, since the 1970s there has been a growing discourse aimed towards undermining the welfare state that situates the poor as an undeserving group that takes advantage of the services provided by taxes paid by middle classes.[372] While the latter is a more obvious form of discrimination, the former helps to reinforce the dehumanisation of poverty.

The fact that these narratives lead to the consideration of people and groups with less economic resources as "others" means that the particular experiences they endure are not considered in the structuring of society. Furthermore, the stigmatisation of poverty leads members of these groups to internalise negative stereotypes about themselves, resulting in feelings of guilt, disempowerment and hopelessness, and further hampering their possibilities of improving their integration in communities and their socioeconomic status.[373]

This construction of a differentiated identity for economically vulnerable individuals favours the appearance of relationships of exploitation and exclusion of the poor.[374] New contractualist and egalitarian liberal theories do, nevertheless, focus on inequality from an economic perspective, considering that a wide array of factors play a role in determining socioeconomic status and thus, in most cases, individuals can not be blamed for their economic well-being. Said theories suggest theoretical and practical approaches to redistributing wealth. However, socioeconomic differences and, more importantly, classism and aporophobia are deeply embedded at the very core of Western capitalist societies, which heavily hinders effective redistribution.[375]

The most basic way to help restructure socioeconomic differences is through the right to education, which is recognised and provided as a public service in Western democracies. However, certain elements, such as the lack of access to other services (for instance, extracurricular activities) and the general environment in which individuals born into lower

[371] BAYÓN, M. C., "La construcción del otro y el discurso de la pobreza: narrativas y experiencias desde la periferia de la ciudad de México", *Revista Mexicana de Ciencias Políticas y Sociales*, vol. 60, No. 223, 2015, p. 372.
[372] EUBANKS, V., *Automating Inequality…*, *cit.*, 2017, p. 38.
[373] BAYÓN, M. C., La construcción del otro y el discurso de la pobreza", *cit.*, 2015, pp. 364-365.
[374] WALKER, A. D. & SMITH, L., "Social class oppression as social exclusion: a relational perspective", in HAMMACK, P. L., *The Oxford Handbook of Social Psychology and Social Justice*, New York, Oxford University Press, 2018, p. 246.
[375] OKUN, A. M., *Equality and Efficiency…*, *cit.*, 2015 (1st ed. 1975), p. 1: "The contrast among American families in living standards and in material wealth reflect a system of rewards and penalties that is intended to encourage effort and channel it into socially productive activity. To the extent that the system succeeds, it generates and efficient economy. But that pursuit of efficiency necessarily creates inequalities."

social classes are raised in, hamper the possibilities they have of climbing up the socioeconomic ladder.[376] In addition, the negative narratives that are, in many cases, developed with regard to public schools, leads middle-classes to take their children to private or semi-private schools so that they are not surrounded and brought up in public school environments.[377] These choices decrease the real equality of opportunities that public education should in theory provide.

As it was stated at the beginning of this section, the role that property (wealth) plays in society, particularly when accessing certain services that are basic for the development of individuals, is a key element that helps reinforce the oppression of lower social strata. For example, with regard to education, the fact that those who do not have access to non-public forms of primary and secondary education generally do not fare as well as those who do,[378] means that, being born into a family with at least medium socioeconomic status (sufficient wealth/property) is almost a necessary precondition to not fall within lower socioeconomic groups when individuals reach adulthood.[379] It is necessary to highlight, at this point ,that average socioeconomic differences in adulthood between individuals who went to public and non-public schools does not result from the latter providing better education but from the differences in background and general environment between public and non-public schools.[380]

The lack of social mobility and low equalisation effect of public education systems increases the pervasiveness of classism. Since those born into medium and high social classes tend to stay in said socioeconomic segment, it is easier to view poorer individuals and groups as "others", thus failing to empathise with their situation and circumstances that have lead to their social standing.

[376] CALERO, J., (dir.), *Desigualdades Socioeconómicas en el Sistema Educativo Español*, Madrid, Secretaría General Técnica del Ministerio de Educación y Ciencia, 2007, p. 175.
[377] FERNÁNDEZ LLERA, R. & MUÑIZ PÉREZ, M., "Colegios concertados y selección de escuela en España: un círculo vicioso", *Presupuesto y Gasto Público*, vol. 67, 2012, pp. 111-113.
[378] CALERO, J. (Dir.), *Op. cit.*, p. 175.
[379] CRAWFORD, C. *et al*., *Family Background and University Success: Differences in Higher Education Access and Outcomes in England*, Oxford, Oxford University Press, 2017, p. 123.
[380] FERNÁNDEZ LLERA, R. & MUÑIZ PÉREZ, M., "Colegios concertados y selección de escuela en España…", *cit.*, 2012, pp. 108-109.

Property also becomes a central element in the perpetuation of low socioeconomic status for those born into lower-income social strata due to the difficulties in accessing credit.[381] This does not only hamper the chances that individuals may have in advancing their socioeconomic position, but can also lead to them resorting to alternative forms of credit that include borderline or directly abusive conditions which will, in turn, further reinforce their pre-existing socioeconomic situation.[382] It is also important to mention the vital role that passing wealth from generation to generation within families through inheritances has in perpetuating social differences.

One of the most significant ways in which socioeconomic oppression has operated is through distorting and discrediting the notion of "class", thereby disregarding the issues that arise from socioeconomic differences. Whereas sex and race are to a certain extent visible characteristics, socioeconomic status can be difficult to identify in certain instances. Additionally, the expansion of meritocratic narratives such as "the American dream" which hold that anyone can achieve a high socioeconomic standing if they work hard enough, leads to the understanding of class as an element that can shift depending on the personal effort of each individual and therefore not a characteristic of their identity that will (in principle) remain constant throughout their lives.[383]

The dilution of working-class consciousness has evolved, hand in hand, with the loss in strength of trade unions. The increase in part-time employment and expansion of the services sector means that workers are much more fragmented and cannot organise as easily.[384] Moreover, vulnerable, lower socioeconomic classes can no longer be identified with the concept of "worker" as an employee in traditional industrial economic sectors. The atomisation of the working-class and subsequent loss of class consciousness hinders intra-group identification, thereby enhancing the development of feelings of antipathy and "otherness" even within the very members of lower socioeconomic strata.

---

[381] LUSTIG, N., ARIAS, O., RIGOLINI, J., "Reducción de la pobreza y crecimiento económico: la doble casualidad", paper presented at the *Seminario 'La Teoría del Desarrollo en los Albores del Siglo XXI'*, 2002, pp. 10-11.
[382] RANKIN, K. N., "A critical geography of poverty finance", *Third World Quarterly,* 2013, pp. 551-552.
[383] THOMAS, V. & AZMITIA, M., "Does class matter? The centrality and meaning of social class identity in emerging adulthood", *Identity,* vol. 14, No. 3, p. 196.
[384] BERNACIAK, M., GUMBRELL-MCCORMICK, R., AND HYMAN, R., "European trade unionism: from crisis to renewal?", European Trade Union Institute, Report No. 133, pp. 53, 83; RICHARDS, A. J. & POLAVIEJA, J., "Trade unions, unemployment and working class fragmentation in Spain", *Instituto Juan March de Estudios e Investigaciones*, 1997, p. 4.

Finally, in line with the intersectional perspective that this dissertation aims to incorporate, the role of property as a central element towards the creation of the individual as an independent being, means that socioeconomic status is interlocked with other narratives of oppression.[385] Women, immigrants and racial minorities are therefore, on average, found in a more vulnerable socioeconomic position than white men. Moreover, the deep connection between classism and racism-xenophobia is made particularly obvious given the difference in general feelings and reactions towards immigrants with a lower socioeconomic status and those with greater income and economic resources, such as elite footballers.

3.5.3.4. Other narratives of oppression

The following pages offer a brief account of other narratives of oppression that are present in society and which, although are not the main focus of this dissertation, can also be and are reproduced by algorithmic decision-making and should therefore also be considered by regulators.

i) *Homophobia and transphobia*

Another of the dominant narratives, which largely results from the construction of gender roles, is that of cis heteronormativity. Heterosexuality has been shaped as the normal sexual orientation and thus as the paradigm of what romantic and sexual relations should be, thereby subordinating homosexuality.[386] Since power relations between men and women are built upon the basis of gender roles, which are, at their baseline, expressed through romantic and sexual relationships, whoever deviates from said norms is excluded and condemned.[387] Throughout history a large number of laws have been passed banning and persecuting homosexuality.[388]While nowadays non-normative sexual orientations are formally tolerated in most Western societies,[389] there are still many countries in which, for example, marriage and adoption laws are not equally applied to heterosexual and same-sex couples and feelings of rejection against homosexual women and men, which are sometimes even expressed by

[385] WALKER, A. D. & SMITH, L., "Social class oppression as social exclusion…", *cit.*, pp. 246-247.
[386] MCRUER, R., *Crip Theory: Cultural Signs of Queernes and Disability*, New York, New York University Press, 2006, p. 7.
[387] WHITLEY JR., B. E., "Gender-role variables and attitudes toward homosexuality", *Sex roles*, vol. 45, No. 11/12, 2001, pp. 701-702; SULLIVAN, M. K., "Homophobia, history, and homosexuality", *Journal of Human Behavior in the Social Environment,* vo. 8, No. 2-3, 2004, p. 2.
[388] SULLIVAN, M. K., "Homophobia, history, and homosexuality", *cit.*, 2004, pp. 1-13.
[389] HEREK, G. M., "Beyond "homophobia": thinking more clearly about stigma, prejudice, and sexual orientation", *American Journal of Orthopsychiatry*, vol. 85, No. 5, pp. 29-31.

political institutions,[390] are still present, [391] thus a deep shift in social norms is still needed in order to ensure the real acceptance and integration of homosexuality.

A very specific element concerning the dominant narrative regarding heteronormativity is the fact that sexual orientation is not an apparent characteristic. In this vein, it is not uncommon to see how this provides grounds for greater apparent tolerance as long as homosexual individuals behave according to their assigned gender roles.[392] Consequently, homosexuality is consciously hidden in many instances, even in Western societies where it is supposedly accepted, with the objective of avoiding exclusion of any sort.[393] This further hinders the possibilities that the particularities and specific needs of the group may be addressed.

Transphobia operates in a similar way to a certain extent since people who do not identify with their biologically assigned sex, do not follow gender norms and are consequently excluded.[394] Furthermore, the discrimination and exclusion suffered by non-cis communities is still deeply entrenched in social and legal norms in Western societies.[395] A particularly problematic issue that arises with regard to trans individuals is the fact that this characteristic of some individuals is still treated in many respects as a pathology.[396]

All in all, the oppression undergone by members of these identity-groups results from the fact that both homosexuality and non-cis identities erode the social fabric constructed through the female-male dichotomy. The division of public/private spheres of activity and complementarity between masculinity and femininity within the institution of the family is no longer applicable in these cases. Once the rigidity of traditional structures becomes especially apparent, conservative groups respond by aiming to further perpetuate these structures, thus hampering the full integration and elimination of the subordiscrimination suffered by non-heterosexual and cis individuals and groups.

---

[390] ŻUK, P., "One leader, one party, one truth: public television under the rule of the populist right in Poland in the pre-election period in 2019, *Javnost – The Public, Journal of the European Institute for Communication and Culture*, vol. 27, No. 3, 2020, pp. 287-307.
[391] ZICK, A., KÜPPER, B. & HÖVERMANN, A., "Intolerance, prejudice and discrimination: A European report", Berlin, Nora Langenbacher, 2011, pp. 64-66.
[392] HUNT, C. J. *et al.,* "Masculine self-presentation and distancing from femininity in gay men: an experimental examination of the role of masculinity threat", *Psychology of Men & Masculinity*, 2016, vol. 17, No. 1, p. 111.
[393] *Ibidem*.
[394] NAGOSHI, J. L. *et al.*, "Gender differences in correlates of homophobia and transphobia", *Sex Roles*, vol. 59, 2008, p. 521.
[395] HOPKINS, P., "Social geography II: islamophobia, transphobia, and sizism", *Progress in Human Geography*, 2019, p. 5.
[396] MACKINNON, R. K., "Pathologising trans people: Exploring the roles of patients and medical personnel", *Theory in Action*, vol. 11, No. 4, 2018, pp. 74-96.

ii) *Religious discrimination, xenophobia, and discrimination based on political beliefs and language*

Other dominant narratives are built upon the axis of religion, resulting in the discrimination of religious minorities.[397] Religious dominant narratives are generally constructed through historical accounts that society internalises. For example, the Spanish account of the *Reconquista*, presents Muslims as "the other", without acknowledging that the land obtained during said process was not naturally assigned to Christians and that most Spanish citizens come from mixed religious lines.[398] This account has reinforced discrimination resulting from prejudices against both, race and religion and has been further strengthened by the narrative that equates Muslim and Arab people in general with terrorists.

Just like with racism and religious discrimination, nationalism is a narrative of conflict built upon the notion of us versus them surrounding the idea of citizenship. Nationalist narratives lead to the appearance of xenophobia, which in combination with racism and feelings of rejection towards poverty[399] produces the exclusion of immigrant communities.[400] Another way in which xenophobia is expressed in Western societies is through reinforcing narratives of conflict between territories within countries.[401]

Similar narratives of conflict are produced by confronting political ideologies that, although within the general spectrum of democracy are accepted to vary, can also result in instances of discrimination in certain scenarios.[402] Furthermore, the assumption and acceptance of liberalism (in its different forms) as the central and guiding ideology in Western democracies leads to the derision and ridicule of those alternatives which do not necessarily prioritise the liberal concept of freedom over equality.[403]

---

[397] SCHLEUTKER, E., "Discrimination against religious minorities", *Journal of Church and State*, 2018, pp. 1-2.
[398] MARTÍN CORRALES, E., "Maurofobia/islamofobia y maurofilia/islamofilia en la España del siglo XXI", *Revista CIDOB d'Afers Internacionals,* No. 66-67, 2004, pp. 39-51; RANA, J., "The story of Islamophobia", *Souls*, vol. 9, No. 2, 2007, p. 154.
[399] Negative feelings towards immigration are generally expressed to a specific type of immigrant who is perceived to lack resources: CORTINA, A., *Aporofobia…*, *cit.*, 2017, pp. 17-22.
[400] CEA D'ANCONA, M. A., "Immigration as a threat: explaining the changing pattern of xenophobia in Spain", *International Immigration & Integration*, vol. 17, 2016, p. 570.
[401] JUNQUERAS, O., "Proximitats genètiques", *Avui*, 27th August 2008. Available on 26th April 2019: https://s.libertaddigital.com/; SERRANO PARTIDA, R., "El desafío catalán y el fin de la transición democrática", *Razón y Palabra*, vol. 22, No. 100, 2018, p. 89.
[402] WESTWOOD, S. J. *et al.*, "The tie that divides: Cross-national evidence of the primacy of partyism", *European Journal of Political Research*, vol. 57, 2018, pp. 333-354.
[403] FERRE, M. M., "Soft repression: ridicule, stigma, and silencing in gender-based movements", in DAVENPORT, C., JOHNSTON, H. & MUELLER, C., (eds.), *Repression and Mobilization*, Minneapolis, University of Minnesota

A final element that must be addressed is language discrimination, which also includes discrimination regarding individuals' accents. This form of discrimination is particularly difficult to identify as it is not widely recognised or included in human rights legal texts. Nonetheless, it is important to briefly point out its existence given the close connection it has with racial, ethnic and class-based discrimination. Language discrimination is constructed by establishing what is the standard form of a language. Anything (and mainly anyone) that deviates from said forms is considered incorrect and therefore socially rejected or looked down on. Language discrimination can take place by marginalising those who speak a certain language, for example, by offering them fewer services or, within a same language, by discriminating against those who speak with "an accent" or speak dialects.[404]

iii)*Narratives of physical*[405] *autonomy*

Since one of the main expressions of freedom is autonomy, individuals who lack physical and psychic autonomy are placed in a subordinated position with regard to fully autonomous individuals.[406]

a. Ableism

Individuals with real or perceived physical and mental abilities that differ from what has been established as the norm have suffered many forms of prejudice and exclusion throughout history. Many narratives have portrayed these people, especially those with different mental abilities as subhuman, even leading, in recent history, to the inclusion of their mass murder as policy in dictatorial regimes.[407] As with other forms of discrimination, narratives so aggressively and openly prejudiced, which in many instances pathologised individuals with diverse abilities have lost power and are considered socially unacceptable but have been substituted by much more subtle forms of oppression and prejudice which lead to their

---

Press, 2005, p. 143: "…consider the term feminazi that Rush Limbaugh coined to ridicule feminists, and that my students report is widely used on campus to attack any woman who stands up for her rights… Ridicule is a decentralized weapon … used to secure power and privilege in and for a wide variety of nonstate institutions."

[404] LIPPI-GREEN, R., *English with an Accent*, London, Routledge, 2nd ed., 2012; TASA FUSTER, V.*, Llengua i Estat: Suïssa i Espanya davant la Diversitat Lingüística*, València, Universitat de València, Servei de Publicacions, 2019.

[405] The use of the word "physical" encompasses both physical and psychic impairments.

[406] HO, A., "The individualist model of autonomy and the challenge of disability", *Journal of Bioethical Enquiry*, vol. 5, No. 2-3, 2008, p. 198; RUDDMAN, D., L., "Shaping the active, autonomous and responsible modern retiree: an analysis of discursive technologies and their links with neo-liberal political rationality", *Ageing & Society*, No. 26, 2006, p. 194

[407] HODGE, N., "Lives worthy of life: the everyday resistance of disabled people", *Journal of Applied Hermeneutics*, 2016, p. 3.

exclusion and discrimination.[408] In addition, individuals that are considered to have below average cognitive abilities are still denied full citizenship in many Western societies. For example, Spain only recognised the right to vote to individuals with cognitive disabilities in 2018.[409]

According to MCRUER[410] liberal capitalism creates an environment of "compulsory able-bodiedness", reinforced by private corporations and public institutions, which stigmatises anyone who does not fit the normalcy standard.[411] The subordinated construction of disability, he argues, results from defining the human body and its functions with regard to capitalist labour.[412] Hence, those who do not fit the normative standard of physical and mental ability do not only suffer discrimination through a general lack of adaptation to their needs, such as architectural barriers, but also see themselves portrayed as inferior, leading to society's and their own internalisation of the structure of subordination.[413]

It is, thus, highly important to point out the differences between the medical and social models of disability. The former has tried to address differences in physical or mental ability from a neutral perspective but approaches said differences as deficiencies.[414] Conversely, the social model establishes a difference between disability and impairment understanding that "a person becomes disabled not because of an impairment […], but because the social and physical environments make living with the impairment challenging."[415] Addressing situations of diverse abilities from the perspective of the social model helps to integrate these individuals and groups as part of the community.

[408] MCRUER, R., *Crip Theory: Cultural Signs of Queernes and Disability*, *cit.*, 2006, p. 2.
[409] Organic Act 2/2018 modifying the General Electoral Regime Organic Act 5/1985 to guarantee the right of suffrage of all persons with disabilities.
[410] MCRUER, R., *Crip Theory: Cultural Signs of Queernes and Disability*, *cit.*, 2006, p. 2.
[411] HARNISH, A., "Ableism and the Trump phenomenon", *Disability & Society*, vol. 32, No.3, 2017, p. 424; MCRUER, R., *Crip Theory: Cultural Signs of Queernes and Disability*, *cit.*, 2006, p. 2.
[412] MCRUER, R., *Crip Theory: Cultural Signs of Queernes and Disability*, *cit.*, 2006, p. 8.
[413] CAMPBELL, F. A. K., "Exploring internalized ableism using critical race theory", *Disability & Society*, vol. 23, No. 2, 2008, p. 152; LOJA, E. *et al.*, "Disability, embodiment and ableism: stories of resistance", *Disability & Society*, vol. 28, No. 2, 2013, p. 194.
[414] GONZÁLEZ RAMS, P., "Las mujeres con discapacidad y sus múltiples desigualdades…", *cit.,* 2010, pp. 2739-2741.
[415] BERRIDGE, C. W. & MARTINSON, M., "Valuing old age without leveraging ableism", *cit.*, 2018, p. 86.

b. Ageism

Ageist narratives are clearly reflected when successful aging is portrayed as not acquiring the characteristics that are typical of old age.[416] The continuous depiction of this ideal version of ageing[417] contributes to the lack of recognition in all aspects of socioeconomic and political life of the specific needs of the elderly.[418]

It is also important to highlight that ageism is not only reflected as discrimination against the elderly but, for example, against older people who are close to retirement in the workplace[419] thereby generating feelings of exclusion early on in the first years of old age.

3.5.4. Incorporating structural discrimination as an element of analysis in the proportionality test

This thesis builds on political and legal critical theories which argue that social structures have been built by members of dominant groups and therefore have not and do not consider the specificities and disadvantaged position of groups that have been historically oppressed.[420] Dominant theoretical narratives, which are currently largely portrayed and originated in RAWLS' thought and derived from his works,[421] have thus become narratives of domination whose partiality is hidden under an appearance of objectivity.[422] These theories, which lie at the cornerstone of the power structures in our society, thus lead to the construction of all social structures, including legal instruments and automated decision-making mechanisms, as apparently neutral elements that in practice perpetuate the subordination of traditionally disadvantaged groups.

This Chapter conveys how the apparent neutrality of liberal doctrines does not address the specific experiences of traditionally oppressed groups, leading to the perpetuation of their

[416] *Ibidem.*
[417] Negative misrepresentations of the elderly which contributed to their open social exclusion have been generally replaced by unrealistically positive portrayals that lead to a much more subtle form of racism. LOOS, E. & IVAN, L., "Visual ageing in the media", in AYALON, L. & TESCH-RÖMER, C. (eds.), *Contemporary Perspectives on Ageism*, Cham, Springer, 2018, p. 170.
[418] BERRIDGE, C. W. & MARTINSON, M., "Valuing old age without leveraging ableism", *cit.*, 2018, p. 85.
[419] NEUMARK, D., BURN, I. & BUTTON, P., "Is it harder for older workers to find jobs?…", *cit.*, 2019, p. 966.
[420] PATEMAN, C. & MILL, C., *Contract & Domination*, *cit.*, 2007, pp. 4-5.
[421] *Idem*, p. 77.
[422] HARAWAY, D., "Situated knowledges: the science question in feminism and the privilege of partial perspective", *Feminist Studies,* vol. 14, No. 3 p. 583: "The moral is simple: only partial perspective promises objective vision. All Western cultural narratives about objectivity are allegories of the ideologies governing the relations of what we call mind and body, distance and responsibility".

disadvantage. More importantly, the previous pages have shown how structures of inequality are deeply pervasive and conform the blueprint upon which social institutions are built. This reality is deeply engrained in the DNA of Western societies, the members of which, either consciously or unconsciously, help to perpetuate the disadvantage of vulnerable groups. The perpetuation of historical oppression and disadvantage through the actions of individuals and the way in which social structures and institutions are shaped is reflected in a wide variety of forms, amongst which algorithms that affect the lives of individuals in a direct or indirect manner can be found.[423]

Hence, the grounding principle that underlies this dissertation that the construction of society through a series of dominant narratives built from liberal perspectives lead to social norms that, while apparently neutral, result in the perpetuation of the disadvantage suffered by members of certain groups. It is therefore necessary to introduce structural discrimination as a category of analysis when balancing out equality and other competing interests. The aim of introducing this notion is not to present equality and non-discrimination as absolute values that must in all cases trump competing principles but to take a step back and re-examine interactions and relationships in the reality of a social context that is still heavily mediated by dominant narratives that place the members of certain groups at a disadvantage.

The existence of a general structural and institutional means that, what may sometimes be presented as a conflict between equality and freedom or efficiency, may in fact not be so if addressed from a different angle. If a step back is taken and the concepts of freedom and efficiency are re-examined, it is possible to find many scenarios in which equality, freedom and efficiency work in a mutually reinforcing manner. These premises must be considered and established as the blueprint for any regulatory or policy instrument aimed towards dealing with instances of discrimination or promoting equality.

Freedom and equality can, thus, also be understood to be working in a mutually reinforcing manner in many cases. Society and power structures are still built upon dominant narrative values that prevent traditionally oppressed groups from really enjoying equal opportunities and thus being free. Consequently, it is not solely a matter of allowing individuals to act freely but also of removing obstacles based on dominant values in order to guarantee that oppressed groups no longer endure structural discrimination and thus enjoy real equal

[423] WINTERS, N., *et al.*, "Can we avoid digital structural violence in future learning systems?", *Learning, Media and Technology*, vol. 45, No. 1, 2020, pp. 17-30.

opportunities and freedom.[424] Similarly, what may in the short-term seem an inefficient measure aimed only towards promoting inequality in many cases ends up proving to bring about more economically efficient results.[425]

Furthermore, even without analysing whether discrimination is efficient or inefficient, it is important to consider that equality and non-discrimination constitute public interests and are ends in themselves. Considering the historical discrimination suffered by the members of disadvantaged groups, and the inherent injustice of the fact that existing structures of discrimination persistently lead to the systemic inequality of historically oppressed groups and of treating individuals differently based on their membership to said groups, advancing equality may be sometimes justified even when it does not generate economic gains.[426] As the World Bank Report "Business, women and the law 2016: getting to equal" stated:

> "We cannot forever remain victims of the idea that the agenda of inclusion and equality (pertaining not just to women but to any group) has to be justified as a means towards the end of higher economic growth. Indeed, what we need to argue is that, even if we had to sacrifice some economic growth in order to achieve inclusion and greater equality, the trade-off would be well worth it. Fortunately, to the best of our knowledge there is no trade-off."[427]

All in all, considering members of groups that have suffered historical oppression are still systematically disadvantaged through the existence of structural discrimination, establishing mechanisms aimed towards protecting the rights to equality and non-discrimination of members of disadvantaged groups is fully justified even though, in some cases, said protection requires limiting the freedom of others.

This chapter has explained the way in which dominant narratives (and narratives of domination) have been historically constructed. The following chapter focuses on analysing

---

[424] GIEBLER, H. & MERKEL, W., "Freedom and equality in democracies: is there a trade-off?", *International Political Science Review*, vol. 37, No. 5, 2016, pp. 602; OKUN, A. M., *Equality and Efficiency…*, *cit.*, 2015 (1st ed. 1975), pp. 23-30.

[425] OKUN, A. M., *Equality and Efficiency…*, *cit.*, 2015 (1st ed. 1975), p. 77: "…what is good for equality may be good for efficiency. THe narrowing of racial differentials during the sixties implied a gain of nearly one-fifthe in the wages and salaries of blacks. That gain approached 1 percent of the nation's income. When we can have more jsutice and more real GNP, society should make the most of it."

[426] ESQUIVEL, V., "Efficiency and gender equality in growth theory: simply add-ons?", *Canadian Journal of Development Studies / Revue canadienne d'études du développement*, vol. 38, No. 4, 2017, p. 550.

[427] IQBAL, S., "Business, women and the law 2016: getting to equal (English)", Washington D.C, World Bank Group, 2015, p.1.

how these narratives permeate the creation and deployment of algorithms used in decision-making processes that affect individuals.

CHAPTER III. HOW MACHINE LEARNING ALGORITHMS AND MODELS CAN DISCRIMINATE

The vast amount of information currently available and the development of technologies that are able to analyse it, generate more efficient and effective decision-making in both public and private organisations.[428] However, this increase in efficiency does not come without a series of errors that can be originated in datasets, in the process of creating algorithms or even appear once systems have been deployed, and which can lead to unfair and discriminatory decisions.[429] As the following pages will convey, there are many moments during the development and deployment stages in which algorithms can become discriminatory and, even when this is not the case, the way in which humans decide to use automated systems can also help to reinforce group disadvantage.

Before delving into how the machine learning model creation process can generate biased and discriminatory decisions it is important to state that all of the tools described in chapter I are inherently discriminatory[430] in the sense that they, for example, choose to measure certain characteristics instead of others or that they determine that some variables have a more important role than others in determining certain results. The discriminatory nature of decision-making systems, understood in the broad sense, is not negative *per se*. Simply choosing between two objects does not necessarily entail unfair discriminatory consequences for certain individuals or groups of people. Nonetheless, as much of the scholarship has pointed out,[431] automated decision-making systems in many cases reinforce pre-existing patterns of subordiscrimination towards traditionally disadvantaged groups. This is the object of study we focus on.

Since this research is developed from a legal perspective, the main focus is set on analysing how existing regulatory instruments can address algorithmic discrimination and on providing a set of regulatory proposals to address the problems caused by discriminatory algorithms. In order to analyse how existing regulations can address the risks and harms generated by algorithms and to put forward new regulatory proposals, it is useful to, at least, gain a general

[428] CUSTERS, B., "Data dilemmas in the information society: introduction and overview", *cit.*, 2013, p. 1.
[429] *Ibidem.*
[430] BAROCAS, S. AND SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 677.
[431] BAROCAS, S. AND SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, pp. 671-732; CITRON, D. K. & PASQUALE, F., "The scored society…", *cit.*, 2014; DATTA, A. *et al.*, "Proxy non-discrimination in data-driven systems: theory and experiments with machine learnt programs", 2017, p. 1. Available on 15th February 2019 at: https://arxiv.org/; O'NEIL, C., *Weapons of Math Destruction…*, *cit.*, 2017.

understanding of the way in which machine learning models and algorithms are created and, more specifically, the ways in which discrimination is embedded in automated decision-making tools. The main reason for this is that by understanding some of the key elements that take part in producing discriminatory algorithms, regulators will be able to develop a comprehensive legal framework that does not only focus on responding to discriminatory algorithmic-generated decisions but that can also prevent discriminatory outcomes from being produced by establishing rules that force data scientists to consciously think about avoiding biases from being embedded into the algorithm during the creation and training process.

Furthermore, being aware of the steps involved in introducing biases in algorithms can also help towards solving the accountability and liability problem that many scholars and public bodies have discussed at large. Knowing the extent to which data scientists' intervention and decisions may condition the discriminatory outcomes produced by the algorithm will provide the necessary elements of proof to make them accountable.

Throwing some insight into "the black box" [432] inside automated decision-making technologies helps to weaken the argument used by some data scientists who consider machine learning and data mining as a form of art,[433] which reinforces the idea that it is impossible to understand and therefore properly control automated decision-making.

Understanding and explaining all the exact processes which take place through algorithmic data analysis and decision-making is not possible given the myriad ways in which machine learning technologies, which are the main focus of this research, can work.[434] Taking that into consideration, the following pages include an overview of the main different steps in which biases can be baked into the model when it is being created during the process of supervised learning. The discriminatory outcomes that may arise once automated systems are already operating, namely, unexpected correlations, will also be addressed.

The steps described below do not necessarily occur in a linear way throughout the process of constructing the model. This is not relevant towards the purposes of this research since the objective is to convey the different possibilities of introducing biases in algorithms.

[432] PASQUALE, F., *The Black Box Society…*, *cit.*, 2015.
[433] BAROCAS, S. AND SELBST, A. D., "Big data's disparate impact", *cit.,* 2016, p. 678; LEHR, D. & OHM, P., "Playing with the data…", *cit.*, 2017, p. 717.
[434] LEHR, D. & OHM, P., "Playing with the data…", *cit.*, 2017, p. 669.

Some of the elements explained below are accompanied by real or figurative examples of how discriminatory outcomes might operate. Said examples are portrayed in a simplistic manner that may lead to the erroneous idea that it is easy to prevent and detect algorithmic discrimination. However, the reality of big data and the different techniques used to process it is much more complex and, especially when dealing with machine learning, also involves the constant self-learning and updating capability of the model even once it is deployed. Consequently, the examples provided must only be considered as an illustrative aid of the processes explained below.

1. INTRODUCING DISCRIMINATION IN THE CONSTRUCTION OF ALGORITHMS

1.1. PROBLEM SPECIFICATION, FEATURE SELECTION AND LABEL DEFINITION: PROXY VARIABLES

1.1.1. Problem specification

The machine learning algorithms that are relevant for the purposes of this research are used in order to make predictions or estimations regarding social phenomena or problems. This means that the system must be told what problem or phenomenon it has to make predictions on and how it should be measured.[435] In this first part of the process, data scientists have to transform general goals into variables that can be measured by the finally resulting model.[436]

In order for the computer to understand what its objectives are, it is necessary for data scientists to establish the target variable, which represents the output information the organisation using the automated decision-making tool is interested in.[437] For example, if a bank wants to work towards improving its general financial situation (general goal) it might want to find out when a client is likely to default in their credit card payments (specific goal), an objective for which the general definition for the target variable would be something similar to: risk of credit card payment default.

[435] *Idem*, pp. 672-673.
[436] *Idem,* p. 673.
[437] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 678.

The target variable is then subdivided into class labels, which identify the different possible values of the target variable.[438] Continuing with our example, possible class labels could be: high risk of default, medium risk of default and low risk of default. While this would constitute a fairly simple example, values for the target variable could also be expressed in a continuous manner, which means they could, for instance, be represented by any numerical value from 0 to 1000, as it happens with credit scores.[439] The former is an example of machine learning used in classification while the latter is an example of machine learning used in regression problems.[440]

In short, in supervised learning, the process of problem specification is carried out by translating an abstract goal into a target variable and identifying the class labels or values that the latter is made up of.[441] The target variable is the conceptualisation of what the algorithm has to predict (probability of credit card payment default) and the class labels constitute the possible results that can come up from the automated decision-making process.[442] The final objective is to teach the model to be able to look for the attributes that are relevant for the target variable each time it is fed data on an individual and provide a value for said target variable.

There are certain target variables that can easily and (almost) unquestionably be divided into class labels.[443] If the goal is for an automated car to detect the difference between a tree, an animal or a pedestrian, there will be one single target variable that can be subdivided into the three class labels: tree, animal and pedestrian, three categories that are objective and do not allow any space for interpretation.[444] In these cases the system can make mistakes but it is not as likely to be biased as when it is measuring predictive goals that do not offer a fixed set of measures.[445] What constitutes high-risk or low-risk of credit card default is determined subjectively.

---

[438] *Ibidem.*
[439] MURPHY, K. P., *Machine Learning…*, *cit.*, p. 8.
[440] LEHR, D. & OHM, P., "Playing with the data…", *cit.*, 2017, p. 673; MURPHY, K. P., *Machine Learning…, cit.,* pp. 5-8.
[441] LEHR, D. & OHM, P., "Playing with the data…", *cit.*, 2017, pp. 673.
[442] *Idem,* pp. 673-675.
[443] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 679.
[444] *Ibidem*.
[445] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 679; LEHR, D. & OHM, P., "Playing with the data…", *cit.*, 2017, p. 674-675.

There are two very closely related steps in the determination of measurements for the target variable that are particularly risky and can help embed biases in the algorithm, especially when predictive goals, that is, target variables, can be measured in multiple ways: feature selection and label definition.[446]

1.1.2. Feature selection

Feature selection is the process by which data scientists choose the attributes that will be observed in the analysis[447] or, in other words, the input variables. Feature selection is connected to the perpetuation of subordiscriminatory structures through encoding proxy or redundant variables for protected group membership. Choosing to measure social phenomena through variables that, while apparently neutral are, in fact, indicative of disadvantaged group membership leads to instances of indirect discrimination.

Many of the attributes included in datasets may seem innocuous and do not directly show a person's inclusion in a group at risk of being discriminated; however, they will very possibly be correlated to protected group membership.[448] Consequently, even when the algorithm is taught not to test attributes such as race or sex, it is very possible that it will continue to learn certain characteristics that are proxies for said specially protected categories.[449] A clear example of this is what happened when "redlining" techniques were used in the US. Through "redlining" banks established particularly risky areas, the inhabitants of which should not be granted loans. Said areas were not only mainly populated by low-income families, but also by ethnic minorities.[450]

A different way in which feature selection can lead to discriminatory outputs is by prioritising variables that benefit dominant and harm disadvantaged groups. For example, if

[446] It is important to highlight that statistical concepts, such variable classification into "qualitative" and "quantitative" are not used because they are not relevant for the purposes of this research. The fact that a social phenomenon is predicted or measured through numerical values, that is, using quantitative variables, does not mean that there is less room for bias. It is the process of deciding how to measure a phenomenon, what features should be measured, what value should be attributed to each feature, etc., that determines biases and discriminatory outputs.
[447] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 688.
[448] CALDERS, T. & ŽLIOBAITĖ, I., "Why unbiased computational processes can lead to discriminative decision procedures" CUSTERS, B. *et al*., (eds.), *Discrimination and privacy in the information society: data mining and profiling in large databases*, Berlin, Springer, 2013, p. 47.
[449] HOUSE OF COMMONS SCIENCE AND TECHNOLOGY COMMITTEE, "Algorithms in decision-making", 2018, p. 21.
[450] HUNT, B., "Redlining", *Encyclopedia of Chicago*, 2005. Available on 20th February 2019 at: http://www.encyclopedia.chicagohistory.org/

job applicants' university ranking position is the first element considered in recruiting processes instead of each applicant's specific abilities, individuals who have not attended well-ranked schools are automatically eliminated from recruiting processes. Consequently, if there are fewer members of disadvantaged groups who attend these universities they will systematically be eliminated from the recruiting process.[451]

1.1.3. Label definition

Feature selection is used in order to define the labels into which the target variables are divided. For example, the target variable "creditworthiness" can be divided into high, medium and low. BAROCAS and SELBST[452] put forward the example of a company that might want to predict which of its employees can be considered to be "good", which is a goal for which there is not one single universal value, meaning the identification of the target variable and its different possible values unavoidably contains some degree of subjectivity which can lead to discriminatory results.[453] When shaping "good" into measurable outcomes there are many different elements that can be considered such as punctuality or the amount of time it takes a given worker to perform a certain task.[454]

For example, the data scientist could decide, probably under the employer's order, to define "good" in relation to past employee performance evaluations carried out by humans. In other words, the features that identify a "good employee" would reflect what in the past evaluators have considered to be a good employee. These evaluations have been proven to contain racial[455] and gender bias[456] in many cases. Consequently, the chosen features and, thus, the way in which the target variable and its possible values are shaped will produce biased consequences for disadvantaged groups.

It is therefore highly important to keep in mind the fact that choices made by data scientists from the very beginning of the designing process are embedded in the final outcomes delivered by the machine learning model.[457] There are different elements that may take part in

[451] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 689.
[452] *Idem*, pp. 678-680.
[453] *Idem*, p. 679.
[454] *Ibidem.*
[455] STAUFFER, J. M. & BUCKLEY, M. R., "The existence and nature of racial bias in supervisory ratings", *Journal of Applied Psychology*, vol. 90, No. 3, 2005, pp. 586-591.
[456] CECCHI-DIMEGLIO, P., "How gender bias corrupts performance reviews, and what to do about it", *Harvard Business Review*, 12th April 2017. Available on 7th April 2019 at: https://hbr.org/
[457] LEHR, D. & OHM, P., "Playing with the data…", *cit.*, 2017, p. 675.

deciding how the target variable and its values will be shaped, such as the data scientist's experience in the field for which it is creating the automated decision-making tool, or the fact that the type of model used conditions the way in which the target variable is shaped, as well as the ease with which certain target variable specifications might be measured.[458]

Hopefully, the potential discriminatory effects and other harms resulting from how problems are specified will also be considered. However, it is possible that these elements are not taken into account, thereby, inadvertently, or perhaps even on purpose, discriminating against especially vulnerable groups of people. In this vein, the lack of diversity in the computer science workforce is one of the elements that must be addressed in order to ensure that machine learning algorithms are not designed from an almost exclusive white-male perspective, which is precisely what leads, in many cases, to the automated discrimination of traditionally oppressed groups.[459] This issue will be discussed more extensively later on.

1.1.4. Example 1: credit scoring

Credit scoring offers a clear example of the way in which the redundant encoding through proxy variables for disadvantaged group membership operates through problem specification, feature selection and label definition, leading to discriminatory results. Credit scoring in theory eliminates the prejudiced attitudes that individual bankers who previously made the decisions on granting loans may have had.[460] Moreover, given that it is supposed to provide an accurate measure of the financial capability of a loan applicant the system benefits both the individual and the creditor seeing as no loans that could put too much strain on the applicant's finances, even leading to a payment default, will be granted.[461]

However, a number of variables are used when determining these scores which can lead to the association between loan decisions and specially protected attributes such as race or gender. For example, postal codes are sometimes used as a variable to determine whether a person should be granted a loan. The system compares individuals' postal codes with data regarding the percentage of loans that have been granted and denied to people living in that same area in the past and data on how many of the applicants who were granted some kind of

[458] *Ibidem.*
[459] CRAWFORD, K., "Artificial Intelligence's white guy problem", *The New York Times*, 25th June 2016. Available on 5th April 2019 at: https://www.nytimes.com/; US EXECUTIVE OFFICE OF THE PRESIDENT, "Artificial intelligence, automation and the economy", *cit.*, 2016, p. 29.
[460] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, p. 145.
[461] BALL, K., "Blacklists and black holes…", *cit.*, 2019, p. 70.

credit in said area defaulted in returning it.[462] If the percentage of credit denials and/or default in that postal area is high, it is less likely that individuals will be granted the loan they apply for.

The problem with using applicants' postal codes is that racial minorities at risk of discrimination, which generally comprehend a significant percentage of immigrant population, tend to live in the same neighbourhoods, in which individuals of a lower socioeconomic status also tend to live. Consequently, whether it is because those in charge of making loan granting decisions in the past were prejudiced towards racial minorities and the percentage of denied loans is higher in minority neighbourhoods or whether it is because racial minorities tend to live in poorer neighbourhoods, due to the use of a proxy variable such as applicants' postal code, they are more likely to be denied loans or to have harsher conditions placed on them.[463] In this sense, AVERY *et al.*[464] found that credit scores predicting a higher probability of loan repayment default tend to correlate with areas with a high presence of minority populations. Consequently, harsher loan conditions correlate with minority populations.[465]

Women are also discriminated in access to credit partly as a consequence of the gender pay gap, which results in women having generally lower scores than men. Even though the overall debt is lower for women than for men, women generally use a higher percentage of their available credit because they have a lower average credit limit, which results in their scores being lowered .[466] Furthermore, HENDERSON *et al.*[467] found that, when using risk scores to determine loan application eligibility for business start-ups, controlling for other factors, the differences in loan conditions due to race and gender were amplified and not reduced.

---

[462] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, p. 146.
[463] JONES HAVARD, C., "'On the take': the black box of credit scoring and mortgage discrimination", *Public interest law journal*, vol. 20, 2011, p. 283.
[464] AVERY, R. B. *et al.,* "Credit scoring: statistical issues and evidence from credit-bureau files", *Real Estate Economics*, vol. 28, 2000, p. 537.
[465] PASQUALE, F., *The Black Box Society…*, *cit.*, 2015, p. 41.
[466] ZARYA, V., "Why being a woman hurts your credit score", *Fortune*, 10th February 2016. Available on 19th April 2019 at: http://fortune.com/
[467] HENDERSON, L. *et al.*, "Credit where credit is due?: race, gender, and discrimination in the credit scores of business startups" *The Review of Black Political Economy*, vol. 42, 2015, p. 477: "…not only do credit scores fail to explain racial and gender differences in credit lines, they appear to mask the size and significance of such differences."

Another recent example of the way in which algorithms discriminate against women in access to credit arose with regard to the differences in credit limit set for men and women in the Apple credit card, which is issued by Goldman Sachs. Several cases in which women who had the same exact economic conditions as their husbands, and even in some instances, had been assigned higher credit scores, where nonetheless set with a much lower credit limit than their male counterparts.[468] However, in this case, it has not been determined where the unequal treatment originated.

One of the main advantages of the new technologies that are in constant development is that, once the fact that it is known that using "postal code" as a variable for determining probability of loan repayment default serves as a proxy for racial minorities and immigrant populations, it is possible to command the algorithm to exclude said variable. However, there are many more variables that may be used and that could also work as proxies for race or other attributes that make an individual at risk of being discriminated. For example, the use of "home ownership" as a variable, which will bring down an individual's credit risk will privilege white men over women and racial minorities.[469]

Moreover, the models used when determining loan eligibility generally use profiling by processing an individual's personal data and comparing it to the information of millions of other people. Consequently, the scores assigned to people who, for instance, have made similar purchases or online searches will influence an individual's score.[470] Additionally, the increasing complexity of algorithmic systems means that the relationships drawn by algorithms between certain variables and personal attributes especially susceptible to discrimination, such as gender or race, will be much harder to detect.[471]

1.1.5. Example 2: health scores

Medical insurance companies want to maximise profit and reduce costs. Doing so by using patients' data in order to draw accurate predictions of health developments will not only benefit the patient but also the company. However, this may also result in discriminatory practices[472] seeing as, the more accurate the information medical insurance companies get, the more likely

[468] TELFORD, T., "Apple Card algorithm sparks gender bias allegations against Goldman Sachs", 11th November 2019. Available on 23rd January 2020 at: https://www.washingtonpost.com/
[469] JONES HAVARD, C., "'On the take'…", *cit.*, 2011, p. 282.
[470] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, p. 145.
[471] See Part I Chapter 2.
[472] MALGIERI, G. & COMANDÉ, G., "Sensitive-by-distance…", *cit.*, 2017a, p. 231.

they will be to incur in price discrimination, setting higher insurance primes for those individuals who, for example, carry out more unhealthy lifestyles,[473] or even, as it has been documented, altogether refuse to insure certain individuals.[474]

Health insurance companies collect data on individuals' purchases,[475] quantified-self apps, financial information and other elements in order to determine their health risk and adapt their insurance premium accordingly.[476] This leads to the discrimination of lower-income segments of the population and minorities seeing as the automated decision-making systems used by health insurers are taught that these populations tend to live in dangerous areas.[477]

In countries in which there is a strong public health system, these health scores should, in theory, not be very problematic since most medical attention is covered by public systems. However, given that, as it was already stated in chapter I, public health is undergoing a process of increased privatisation,[478] and private insurance is already necessary in many countries, discriminating against already vulnerable groups will mean further deepening situations of inequality with regard to a basic need such as healthcare.

1.1.6. Example 3: predatory advertising

Predatory practices in advertising create profiles of consumers with the objective of targeting especially vulnerable groups of people with advertisements of low quality or toxic products and services.[479] This example is particularly relevant as it shows how algorithms and profiling techniques are used to purposefully disadvantage vulnerable populations.

The final decision regarding the purchase of a certain product or service is in the consumer's hands. However, when a firm's target audience is made up of especially vulnerable people,

[473] MARWICK, A. E., "How your data are being deeply mined", *New York Review of Books*, 9th January 2014. Available on 20th February 2019 at: http://www.tiara.org/
[474] PASQUALE, F., *The Black Box Society…*, *cit.*, 2015, pp. 26-27.
[475] BECKETT, L., "Everything we know about what data brokers know about you", *Propublica*, 13th June 2014: "One health insurance company recently bought data on more than three million people's consumer purchases in order to flag health-related actions, like purchasing plus-sized clothing, the Wall Street Journal reported. (The company bought purchasing information for current plan members, not as part of screening people for potential coverage.)"
[476] US EXECUTIVE OFFICE OF THE PRESIDENT, "Big data and differential pricing", 2015, p.7.
[477] ALLEN, M., "Health insurers are vacuuming up details about you — and it could raise your rates", *Propublica*, 17th July 2018.
[478] SÁNCHEZ-MARTÍNEZ, F. I., ABELLÁN-PERPIÑÁN, J. M. & OLIVA-MORENO, J. "Privatization in healthcare management…", *cit.*, 2014, pp. 75-80.
[479] US SENATE COMMITTEE ON COMMERCE, SCIENCE & TRANSPORTATION, MAJORITY STAFF, "A review of the data broker industry: collection, use, and sale of consumer data for marketing purposes", 18th December 2013, p. i.

the quality of the product they are selling as well as the 'freedom' with which consumers make the final purchasing decision should be put into question. The labels used by firms, such as the ones pointed out in the following paragraph, are clearly designed to attract a subset of the consumer population which has less resources and is therefore less capable of comparing different product offers and detecting abusive conditions in contracts.

A report published by the US Senate Committee on Commerce, Science & Transportation in 2013 found that some of the labels used in consumer profiling were specifically designed to point out especially vulnerable consumers. Some of the categories included in the report were: "'Rural and Barely Making It,' 'Ethnic Second-City Strugglers,' 'Retiring on Empty: Singles,' 'Tough Start: Young Single Parents,' and 'Credit Crunched: City Families.'"[480] An especially worrying element of this example is that this is a case in which discrimination was purposefully included in the algorithm through the labels associated to the individuals in the sample.

Another report published in 2012 by the US Senate on for-profit colleges shows that the target groups training recruiters were taught were: "Welfare Mom w/Kids. Pregnant Ladies. Recent Divorce. Low Self-Esteem. Low Income Jobs. Experienced a Recent Death. Physically/Mentally Abused. Recent Incarceration. Drug Rehabilitation. Dead-End Jobs-No Future."[481] The vast amount of information that firms using predatory advertising can now collect in combination with new machine and deep learning technologies enhances their chances at reaching many more vulnerable people. These companies can now feed their marketing algorithms with massive amounts of data so that the automated tools they use are automatically able to detect many more prospective consumers than the methods they previously employed.

1.2. DATA COLLECTION

"Actual data" consists of known information regarding an individual while "modelled data" is the output information reached through inferences carried out by the model based on the "actual data".[482] During the training phase, machine learning algorithms and models learn

[480] *Idem*, p. ii.
[481] US SENATE HEALTH, EDUCATION, LABOR AND PENSIONS COMMITTEE, "For profit higher education: the failure to safeguard the federal investment and ensure student success", 2012, p. 58.
[482] US SENATE COMMITTEE ON COMMERCE, SCIENCE & TRANSPORTATION, MAJORITY STAFF, "A review of the data broker industry…", *cit.*, 2013, p. 22.

from past examples, also known as training data. The training data are the data fed into the computer programme in order for it to extract the relevant relationships using data mining and/or machine learning techniques.[483] The actual data needs to be collected and prepared before it is processed.

Given that the training data are key in constructing the logic or rules according to which the model will work once it is deployed, the quality of the data collected must be good.[484] However, big data analysis tools such as data mining or machine learning provide better results when there is more data available,[485] which means that that a trade-off between clean data and efficiency must be accepted.[486]

The elements leading to the use of biased training datasets listed below are assumed to occur unintentionally, however, it is also possible to find cases in which sample selection is intentionally biased.[487]

1.2.1. Unrepresentativeness in the dataset

Data scientists must ensure that the data collected is representative of the population[488] since, if this is not the case, biased samples will lead to data mining and/or machine learning drawing inferences from them as if they were actually representative of the population and, consequently, the results generated will systematically place those who are under or, in some cases, overrepresented in the sample in a disadvantaged position when decisions are made based on the inferences carried out by the algorithm.[489]

The lack of representativeness in the sample used may occur, for example, as a consequence of inadvertently excluding some sectors of society. People who live in the margins of datafication, especially due to lack of resources, can undergo further marginalisation if they are systemically omitted from the training data samples.[490] If the trend to increasingly base

[483] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 680.
[484] SURDEN, H., "Machine learning and law", *cit.*, 2014, p. 106.
[485] LEHR, D. & OHM, P., "Playing with the data…", *cit.*, 2017, p. 678.
[486] CUKIER, K., and MAYER-SCHOENBERGER, V., "The rise of big data…", *cit.*, 2013, p. 29.
[487] CALDERS, T. & ŽLIOBAITĖ, I., "Why unbiased computational processes can lead to discriminative decision procedures", *cit.*, 2013, p. 47.
[488] ŽLIOBAITĖ, I., "A survey on measuring indirect discrimination in machine learning", September 2015. Available on 25th September 2018 at: https://arxiv.org/
[489] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 681.
[490] LERMAN, J., "Big data and its exclusions", *cit.*, 2013, p. 57; CRAWFORD, K., "Think again: big data", *Foreign Policy*, 10th May 2013. Available on 25th September 2018 at: https://foreignpolicy.com/

decisions on the information obtained from big data analysis techniques continues to grow, some sectors of society could be completely excluded from having their preferences and needs considered by businesses and governments thus hampering the access to certain services of those who most need them.[491]

In other cases, the underrepresentation of a certain group of people in the training data is derived not from their lack of access to technology, but from past cases of discrimination. Take for example the case of promotions. If historical data on promotions is fed into the computer programme designed to indicate what characteristics are more common in employees who have been promoted in the past, most of the subjects appearing in the dataset will be male.[492]

A clear example of this is Amazon's recruiting algorithm, which was discarded after the team in charge of testing it, concluded that it was sexist.[493] The tool, which employed machine learning technology, was learning from the information of CVs submitted to Amazon over the previous ten years.[494] Given the fact that employment in technology is male-dominated, most of the CVs submitted and people hired had been men. Consequently, the algorithm learned that being male was preferable and thus punished CVs containing the word women and candidates who, for example had attended all women's colleges.[495] It is not clear whether the algorithm was also biased because past hiring decisions had discriminated against female applicants or whether it was solely due to the fact that there are more men than women in STEM careers, however, it has been assumed that it was due to the latter. In any case, its discriminatory effects are certainly unquestionable.

Conversely, it is also possible that overrepresentation of a particular group may lead to situations of discrimination. For example, say in any given city there has been a higher degree of arrests in a certain neighbourhood due to the fact that it is mostly populated by racial minorities, who are more frequently targeted by the police. The disproportionate amount of arrests of people belonging to a racial minority would probably lead to, for example, black and Hispanic people who have committed crimes, being caught at a much higher frequency than white criminals. Consequently, if the subjects whose data appears in

[491] LERMAN, J., "Big data and its exclusions", *cit.*, 2013, pp. 57-58.
[492] DATTA, A. *et al.*, "Proxy non-discrimination in data-driven systems…", *cit.*, 2017, p. 1.
[493] DASTIN, J., "Amazon scraps secret AI recruiting tool that showed bias against women", *Reuters,* 10th October 2018. Available on 12th February 2019 at: https://www.reuters.com/
[494] *Ibidem*.
[495] *Ibid*.

the training dataset are exclusively individuals who have been previously arrested or convicted, when the data is introduced into the algorithm, it will extract a relevant relationship between being black or Hispanic and committing a crime.[496]

This is exactly the problem that has arisen with regard to certain predictive policing tools. One of the leading firms which public law enforcement authorities from all over the world hire in order to introduce these tools in their police forces is PredPol.[497] There are, however, a series of different predictive policing tools which are used by police departments and which all work more or less under the same logic.[498]

These programmes indicate where police officers should be placed, increasing the number of officers in those areas which are considered to be crime hotspots, that is, where more criminal activity is likely to occur.[499] Moreover, these programmes are even able to predict with a certain degree of accuracy the type of crime and place where it will occur[500] as well as when a crime wave is about to happen.[501] The way these tools are designed is based on the techniques used for earthquake prediction, incorporating data from past crimes into its database in order to predict when and where criminal offences are likely to occur.[502]

One of the positive aspects of predictive policing models is that they do not lead to the identification or tracking down of individuals and the main input variables are simply the times and places in which past offences have happened.[503] The system helps to distribute police officers accordingly and thus offer more protection where there is a higher tendency for crimes to be committed. These software programmes do not use personal identifiable information and are, in theory, blind to race, social class, gender and any other element that could lead to profiling individuals as potential criminals because of correlations established with their membership to a specially protected group.[504]

---

[496] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 681.
[497] See PREDPOL. Available, on 14th February 2019 at: https://www.predpol.com
[498] EUROCOP, "Análisis y prevención del delito", 2015. Available on 11th April 2019 at: https://www.eurocop.com/
[499] BENBOUZID, B., "Des crimes et des séismes : la police prédictive entre science, technique et divination", *La Découverte*, No. 206, 2017, p. 101.
[500] EUROCOP, "Análisis y prevención del delito", *cit.*, 2015.
[501] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, p. 85.
[502] BENNET MOSES, L. & CHAN, J., "Algorithmic prediction in policing: assumptions, evaluation, and accountability", *Policing and Society*, vol. 28, No. 7, 2018, p. 808.
[503] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, p. 86.
[504] ABATE, M., "How to prevent crimes using earthquakes", in EMMER, M., & ABATE, M., (eds.), *Imagine Math 6: Between Culture and Mathematics*", Springer, Switzerland, p. 105.

The programme offers police departments two types of crime prediction packages. The first only covers very serious crimes such as murder or large-scale drug trafficking. There is however a second possibility, which allows law enforcement departments to expand the types of crimes detected by the predictive software in order to include lesser offences such as petty theft or selling small quantities of drugs.[505]

The problem with lesser crimes is that they tend to occur in largely impoverished neighbourhoods and, consequently, when these data are introduced into the model, the software tool will be more likely to send police officers to marginalised neighbourhoods in which they will probably arrest more people.[506] Hence, the feedback received by the learning model will reassure it of the fact that more crimes tend to occur in said impoverished areas and the vicious circle of arrests and perceived dangerous areas will continue.[507] These biased results do not only lead to the constant discrimination of people from lower-income backgrounds, who are continuously targeted as potential criminals, but also, and more specifically, to the discrimination of the ethnic minorities who generally live in said neighbourhoods.[508]

Even though the objective of predictive policing tools is to prevent serious crime, police forces sometimes choose to include all the data available due the extended idea that small crimes in neighbourhoods drive law-abiding citizens away thereby creating an atmosphere prone to more serious crime.[509] Thus, by combatting lesser offences and fixing up neighbourhoods, it is in theory possible to also prevent more serious crime.[510] One of the variations of this idea was crystallised through the zero-tolerance policy put in place by the city of New York during the decade of 1990, with the aid of another predictive policing tool, Compstat,[511] which is still currently being used.[512]

Given that when the data for lesser crime is also introduced, the model is able to accurately predict many more offences (most of which are, once again, lesser crimes), it looks like the

[505] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, p. 86.
[506] *Ibidem*.
[507] *Ibid*.
[508] *Idem,* p. 87.
[509] *Ibidem*.
[510] KELLING, G. L. & WILSON J. Q., "Broken windows: the police and neighborhood safety", *Atlantic Monthly*, March 1982. Available on 14th February 2019 at: https://www.theatlantic.com/
[511] GREENE, J., "Zero tolerance: a case study of police policies and practices in New York City", *Crime and Delinquency*, vol. 45, No. 2, 1999, pp. 171-187.
[512] BYFIELD, N. P., "Race science and surveillance: police as the new race scientists", *Social Identities: Journal for the Study of Race, Nation and Culture*, vol. 25, No. 1, 2018, p. 101.

more data on all sorts of crime the predictive software is fed, the better results it will provide.[513] The result is a self-fulfilling prophecy which criminalises the poor as well as racial minorities. This example illustrates a scenario in which algorithms discriminate not only due to learning biases, but also as a result of part of the data introduced being proxy variables for protected group membership.

For this reason, the city of Oakland in the US decided not to use this type of predictive policing tool[514] and some private firms that have reduced the significance that lesser felonies, such as drug-related crimes, have in their models.[515]

1.2.2. Errors in the dataset

Another problem that might lead to biases in automated decision-making processes is the inclusion of errors regarding certain people's data that might lead to discriminatory results. These errors lead to inaccurate depictions of people belonging to disadvantaged groups. One of the reasons why this happens is due to the fact that information is expensive and, consequently, organisations tend to prefer to use cheaper and less accurate datasets.[516] The information and results offered by these datasets is still effective for the purposes of organisations but the reduction in granularity can lead to making erroneous generalisations that can have significant impacts on individuals, especially if they are members or are associated to members of disadvantaged groups.

If inaccurate information is evenly distributed amongst all the subjects included in a dataset, errors should not be a problem and should not produce discriminatory results. However, research has shown that it is more likely for errors to appear in the data of people belonging to traditionally oppressed groups, thereby leading to the systematic production of biased outcomes trained on said errors.[517] For example, problems have arisen with regard to identity verification programmes, which have been found to work with databases that contain a

[513] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, p. 89.
[514] THOMAS, E., "Why Oakland police turned down predictive policing", *VICE*, 28th December 2016. Available on 2nd April 2020 at: https://www.vice.com/
[515] HUNCHLAB, "A citizen's guide to HunchLab", 11th July 2017, p. 26.
[516] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 689.
[517] KIM, P. T., "Data-driven discrimination at work", *William & Mary Law Review*, vol. 58, 2017, pp. 885-886.

greater degree of errors in the data of married women who have changed their surname and individuals with two surnames (mainly Hispanics).[518]

Another example that must be brought forward are the errors found in the datasets used in determining credit scores in the US. The US Federal Trade Commission found that around "five per cent of consumers had errors on their credit reports that could result in less favourable terms for loans."[519] The implications that algorithmic errors may have in perpetuating pre-existing situations of disadvantage are very extensive, particularly given the fact that individuals who belong to vulnerable groups such as immigrant populations generally have less resources and knowledge to challenge certain automated decisions, such as the ones that are used when providing credit scores.

1.3. LABELING EXAMPLES IN THE TRAINING DATASET

In supervised machine learning, the algorithms and models are trained with examples that are labelled according to the different class labels or values that have been established for the target variable.[520] In some cases the examples are already labelled. For example, if a firm decides to use past hiring data in order to determine what characteristics make up a good job candidate and thus be able to predict which applicants would be better suited for the job, the labels would differentiate between hired/not-hired. It is essential to keep in mind that this is not a problem of unrepresentative datasets but of prior conscious or unconscious prejudice that could be introduced into the algorithm if, for example, past hiring decisions had disfavoured women or racial minorities. The difference with the Amazon recruitment algorithm is that, in the case of Amazon we assume past hiring choices were not discriminatory: there were simply less female applicants.

A real case in which biased labels led to discriminatory outcomes took place in St. George's Hospital Medical School which, in 1988, was found guilty by the UK's Commission for Racial Equality (currently the Equality and Human Rights Commission) of discriminating against women and racial minorities in its admissions process. The medical school had been using a computer programme that had automatized its initial selection phase. Since the

[518] US EXECUTIVE OFFICE OF THE PRESIDENT, "Big data…", *cit.*, 2014, p. 52.
[519] US FEDERAL TRADE COMMISSION, "In FTC study, five percent of consumers had errors on their credit reports that could result in less favorable terms for loans", February 11th 2013. Available on April 9th 2019 at: https://www.ftc.gov/news-events/press-releases/2013/02/ftc-study-five-percent-consumers-had-errors-their-credit-reports
[520] See II.1.

programme was based on previous admissions decisions made by staff it simply introduced prior bias into its decisions, thereby rejecting at a much higher frequency female candidates as well as candidates with non-European sounding names.[521] While the example provided took place more than thirty years ago, considering the degree of development that automated decision-making processes have achieved and the many stages and possibilities for introducing biases in algorithms that currently exist, it is highly likely that prior prejudice will be embedded into these decision-making systems.

Other particularly problematic cases of biased labels can take place when the examples are manually labelled by the data scientist.[522] There are different problems that may arise during this process and which mean that a great degree of arbitrariness can be introduced into the algorithm.[523] Firstly, it will probably not be clear where to set the thresholds for the different values or class labels, meaning that the data scientist will decide which subjects fit into which labels or values.[524] These thresholds are not natural but set by humans, which means that they can be biased.[525]

Additionally, even if at the time in which the algorithm is designed there is a general agreement for a threshold being adequate, this might change over time.[526] For example, in a situation of economic crisis it could be agreed that failing to make loan return payments for three consecutive months would automatically introduce the data subject into the low creditworthiness class label. However, if the economic situation changes banks may become more lenient and thus consider that low creditworthiness would result from failure to pay for four consecutive months. While this is a simple example that would in theory pose no modification problems, the interested parties would still have to take the time to introduce the corresponding modifications in the already deployed model. Moreover, when many different pieces of data are combined within each class label it will be very difficult to redefine the values once the model has been deployed.[527]

[521] LOWRY, S. & MACPHERSON, G., "A blot on the profession", *cit.*, 1988, p. 657.
[522] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 681.
[523] HAND, D. J., "Classifier technology and the illusion of progress", *Statistical Science,* vol. 21, No. 1, 2006, pp. 10-11.
[524] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 681.
[525] HAND, D. J., "Classifier technology and the illusion of progress", *cit.*, 2006, p. 11.
[526] CALDERS, T. & ŽLIOBAITĖ, I., "Why unbiased computational processes can lead to discriminative decision procedures", *cit.*, 2013, p. 47.
[527] HAND, D. J., "Classifier technology and the Illusion of progress", *cit.*, 2006, p. 10.

Secondly, some cases may not fit perfectly into one class label or value for the target variable or class labels may not be precise enough to provide a clear picture of the differences between examples.[528] In this scenario, data scientists will have to choose which label or value to assign to non-clear cases, introducing another element of arbitrariness into the equation.

Even if the examples are clearly mislabelled or, even if correctly labelled, represent some sort of discriminatory bias towards especially vulnerable groups of people, given that this is the data that the model has been trained on, when the model is tested against a different set of data it will still generally seem to be making accurate predictions.[529]

1.4. DATA PRE-PROCESSING TECHNIQUES

1.4.1. Missing value imputation

Even though efforts should be made during the data collection process in order to ensure that the datasets used have a certain degree of accuracy, it is very difficult to find datasets that do not contain missing or inaccurate values.[530] For example, in a dataset consisting of surveys some of the respondents might have chosen to not fill out their sex or their age or the person in charge of introducing in a computer the information on handwritten forms can not understand or misreads some of the handwritten information.

While a possible solution would be to simply delete the subjects with missing values from the dataset, the fact that machine learning needs very large quantities of information in order to work properly means that this alternative will not always be viable.[531] Some algorithms have been developed in order to be able to impute missing values by comparing the rest of the information in the data subject's profile with other subjects and thereby attributing the data subject the same value as the one held by similar individuals.[532] This, however, could lead to imputing wrong values[533] and to profiling in the cases in which specially protected group membership was the imputed value or used in order to impute another value.

---

[528] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 681.
[529] HAND, D. J., "Classifier technology and the Illusion of progress", *cit.*, 2006, p. 10; BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 682.
[530] LEHR, D. & OHM, P., "Playing with the data…", *cit.*, 2017, p. 681.
[531] *Ibidem.*
[532] BREIMAN, L. & CUTLER, A., "Breiman and Cutler's Random Forests for Classification and Regression", 2018, p. 10. Available on April 5th 2019 at: https://cran.r-project.org/
[533] LEHR, D. & OHM, P., "Playing with the data…", *cit.*, 2017, p. 682.

1.4.2. Dimensionality reduction and feature extraction and construction

Data pre-processing techniques can be very useful in order to clean the data and extract the relevant information from the sample. There are several mechanisms, such as value imputation, which was already explained, that can be employed so that when the data is fed into the algorithm it is cleaner and easier to process.[534] For example, feature extraction, selection and construction are used in order to reduce the amount of data that will be fed into the computer so that it is more manageable and once the computer processes it, the results obtained are more accurate.[535]

Through feature extraction and construction, a certain attribute, date of birth, could be transformed into age if it was considered to be more relevant or better for the analysis. Different attributes can also be combined to form a different feature that will be measured, for example, height and weight could be combined and the actual feature analysed by the algorithm would be body mass-index.[536]

Another technique is dimensionality reduction, by which different elements are grouped into one category in order to reduce redundancy. For example, for a spam-filtering algorithm, dimensionality reduction could be done by, instead of telling the algorithm exactly every word that it needs to check for in e-mails, to teach it that e-mails containing many medical-related terms should be considered to be spam.[537]

1.5. DIVIDING THE DATASET

Models are trained by learning from a set of data and then by being fed a second set of data in order to test its predictive ability. If the training dataset is larger, the model will learn to make better predictions. However, if the test set is too small, the data scientist will be less aware of the extent to which the model's predictions are actually generalizable for real world data.[538]

---

[534] LIU, H. & MOTODA, H., "Feature selection, extraction and construction", paper presented at the *Towards the Foundation of Data Mining Workshop, Sixth Pacific – Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2002)*, Taipei, Taiwan, 2002, p. 67.
[535] LIU, H., AND MOTODA, H., "Feature Selection, extraction and construction", *cit.*, 2002, p. 67.
[536] CALDERS, T. & CUSTERS, B., "What is data mining and how does it work?", in CUSTERS, B., *et al.*, (eds.), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in large Databases*, Berlin, Springer, 2013, p. 39.
[537] *Ibidem*.
[538] LEHR, D. & OHM, P., "Playing with the data…", *cit.,* 2017, p. 686.

1.6. MODEL SELECTION

A very important part of the model creation process is selecting the actual type of model that will be used. This choice will determine the type of mechanisms used when generating predictions. Drawing from LEHR and OHM's[539] work, the following paragraphs contain a brief explanation of some model characteristics that may take part in the algorithm disfavouring members of disadvantaged.

Firstly, there are certain types of models (when the possible outcome is binary) that are more prone to false negatives than to false positives.[540] Closely related to this, a choice can also be made in models with continuous value outcomes, between a tendency to overestimate or underestimate.[541] Consequently, for example, in cases of models designed for their use in the criminal justice system, a recidivism risk model more prone to false negatives or to the underestimation of recidivism risk could be prioritised given the implications that these tools can have for the fundamental rights of individuals and other core values of democratic states.[542]

Another element that can be chosen is the extent to which the model is more or less vulnerable to "overfitting". Overfitting means that algorithms tend to look for relationships in the data to the extent that it will sometimes find correlations even when different pieces data have nothing to do with each other.[543] This happens because when the system is deployed it continues to trust certain relationships that were relevant during the training and testing stages but that lose their significance once the algorithm has been deployed.[544]

Choosing a model with less tendency or vulnerability to overfitting will, in theory, provide more accurate results and reduce the risk for discriminatory outcomes as unexpected

[539] *Idem*, p. 688-695.
[540] *Idem*, p. 656-657.
[541] *Idem*, p. 691.
[542] *Idem*, p. 656-657.
[543] *Idem*, p. 684.
[544] SCATAMBURLO, T., CHARLESWORTH, A. & CRISTIANINI, N., "Machine decisions and human consequences", "Machine decisions and human consequences", in YEUNG, K. & LODGE, M., (Eds.), *Algorithmic Regulation*, Oxford, Oxford University Press, 2019, p. 57.

correlations that work by punishing historically disadvantaged groups will be less likely to come up.[545]

One of the most relevant issues that arise with regard to automated decision-making is the lack of transparency of these new technologies and the difficulties in getting to know the logic underlying the models used.[546] Although the degree to which the model is or not explainable does not directly influence possible discriminatory outcomes, choosing a model with a higher degree of explainability will facilitate controlling the results it produces.[547] Additionally, being aware of the fact that a higher degree of accountability is possible could also serve as a deterrent for data scientists tempted to encode some sort of bias into the model.

2. ALGORITHMIC DISCRIMINATION THROUGH CORRELATIONS: FAULTY AND PRECISE INFERENCES

Once algorithms have been deployed, unexpected correlations can come up.[548] These correlations can have negative consequences for disadvantaged groups when sensitive attributes such as race, sex or socioeconomic status are explicitly used or when other variables that act as proxies for vulnerable group membership operate. The correlations established between different pieces of data which lead to negative outcomes for members of disadvantaged groups can be the result of the way in which the algorithm is built but also of what the system learns once it is deployed.

While correlations do not necessarily entail causation, since algorithms are capable of processing such large amounts of data, the results they obtain just by establishing patterns regarding the information they are fed are more accurate than those generated by traditional statistical methods.[549] This increased accuracy does not prevent algorithms from making

[545] LEHR, D. & OHM, P., "Playing with the data…", *cit.,* 2017, p. 704: "Previous scholarship has documented how data, particularly survey data, can often be noisier for minority groups than for others, and an algorithm that overfits risks improperly capitalizing on this noise more so than an algorithm that does not overfit"
[546] PASQUALE, F., *The Black Box Society…*, *cit.*, 2015; CITRON, D. K. & PASQUALE, F., "The scored society…", *cit.*, 2014; KROLL, J. *et al.*, "Accountable algorithms", *cit*., 2017.
[547] LEHR, D. & OHM, P., "Playing with the data…", *cit.*, 2017, p. 692.
[548] CITRON, D. K. & PASQUALE, F., "The scored society…", *cit.*, 2014, p. 5.
[549] ANDERSON, C., "The end of theory: the data deluge makes scientific inquiry obsolete", *Wired*, June 23rd 2008. Available on February 22nd 2019 at: https://www.wired.com/2008/06/pb-theory/: "Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence) […] But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete […] There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses

mistakes as the use of correlations to establish patterns and predict behaviour may also mean that, in some cases, the correlations drawn between two pieces of data will actually have no significance.[550] This is one of the main problems that the scholarship has brought up when addressing discriminatory decisions resulting from algorithms, not only regarding individuals belonging to groups at risk of being discriminated, but also in general regarding unfair decisions, given the fact that automated decision-making systems may draw results from faulty inferences.[551]

Correlations which result in instances of algorithmic discrimination can result from the direct use of sensitive data by the algorithm or from the use of proxies that are associated with protected group membership.[552] The use of sensitive data or proxies for said data acts in combination with the other elements that were previously discussed and which can make the algorithm biased against said groups. For example, if an automated recruiting system is fed with data from previous hiring decisions, even if the actual sex, socioeconomic status or race of job applicants is not indicated it will be possible to extract said information from other data points such as the schools and extracurricular activities they attended. Thus, if the algorithm establishes a correlation between attending a certain school district (which acts as a proxy for race) and not having been hired, it will reproduce said decisions thereby leading to discriminatory outcomes against racial minorities.

Another related, and perhaps even deeper, problem takes place when the inferences carried out by algorithms are actually accurate.[553] For example, it might be true than black men are more likely to have gone to prison than white men. When correlations drawn by algorithms reflect the reality of a disadvantaged group questions will be raised on whether using these methods is legitimate or whether notions of equality should be prioritised given that these accurate correlations generally result from the structural discrimination that members of

about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."

[550] LEHR, D. & OHM, P., "Playing with the data…", *cit.*, 2017, p. 684.

[551] BAROCAS, S., "Data mining and the discourse on discrimination", *Proceedings of the Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining (KDD)*, 2014, p. 2. Available, on 20th February 2019 at: https://pdfs.semanticscholar.org/; MITTELSTADT, B. D. *et al.*, "The ethics of algorithms…", *cit.*, 2016, p. 5.

[552] ZARSKY, T., "Understanding discrimination in the scored society", *Washington Law Review*, vol. 89, No. 4, 2014, p. 1389.

[553] BAROCAS, S., *Op. cit.*, pp. 2-3.

disadvantaged groups have historically suffer and are still subjected to.[554] The issue of accurate discrimination is addressed to a larger extent in the following chapter.

3. THE SOCIAL (AND PERSONAL) ORIGIN OF DISCRIMINATORY ALGORITHMS

3.1. PRIOR AND ON-GOING BIASES

In order to be fully aware of the extent to which algorithms can discriminate it is important to determine where biases contained in the algorithm may originate. The origin of discriminatory results may not always be relevant towards their legal treatment seeing as conscious discrimination will be very difficult to prove and will be easily masked by the data and algorithms.[555] However, pointing out the way in which biases in algorithms may be produced draws attention to the dangers automated decision-making systems create and to the need for a regulatory framework.

As some of the examples described prove, algorithm and model design can lead to automated decision-making systems learning from data that is prejudiced given the fact that it reflects biased decisions that were made in the past by humans.[556] A clear example of this is the algorithm developed and used by St George's Hospital Medical School.

Algorithms can, however, also reflect on-going biases. As the previous chapter analysed, the prevalence of formal equality as a valid form of real equal opportunities leads to the persistence of structural discrimination. Said structural or systemic discrimination means that disadvantaged groups continue to have fewer political and economic resources and, consequently, fewer opportunities and that certain conscious or unconscious prejudices against individuals whose identity belongs to one or more of said groups are still held. This leads to decisions that continue discriminating against especially protected groups and which, when fed to algorithms, result in initially neutral algorithms learning from on-going biases present in society.

[554] This discussion will be addressed in Part III, when formulating the application of the anti-discrimination legal framework to algorithmic discrimination.
[555] BAROCAS, S. AND SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 692-693.
[556] *Idem*, p. 682.

3.1.1. Example 1: Algorithms used in healthcare

Discrimination in healthcare has also been pointed out as one of the risks derived from the widespread use of algorithms. Machine learning technologies may lead to delivering worse healthcare to women and racial minorities seeing as they base models on white men.[557] The gender and racial bias in healthcare and medical research has been widely documented[558] and if machine learning technologies used in healthcare learn from existing and past biases, the gap in the development and quality of medical attention received by white males will widen with respect to healthcare delivered to women and racial minorities.[559]

3.1.2. Example 2: Recidivism risk prediction

The use of recidivism models is very common in the United States.[560] These models determine the likelihood that a convicted offender will engage again in the future in some sort of criminal activity and consequently provide courts with a theoretically valid recidivism risk score that can influence sentencing decisions.[561] As with many of the technologies discussed here, they are in theory blind to circumstances, such as race or social background, which can lead to biases when determining the likelihood of recidivism.[562]

While these models began operating as a consequence of the general consensus that they would come up with less prejudiced results than human beings,[563] as some specific cases have shown, models which process individuals in order to determine their recidivism risk score tend to punish racial minorities and the poor.[564] One of the algorithms used by many US courts was found to predict black defendants would reoffend twice as much as it

[557] GIANFRANCESCO, M. A. *et al.*, "Potential biases in machine learning algorithms using electronic health record data", *JAMA Internal Medicine*, vol. 178, No. 11, p. 4.
[558] DOYAL, L., "Sex, gender, and health: the need for a new approach", *British Medical Journal*, vol. 323, 2001, pp. 1061-1063; NELSON, A., "Unequal treatment: confronting racial and ethnic disparities in health care", *Journal of the National Medical Association*, vol. 94, No. 8, 2002, pp. 666-668; RUIZ-CANTERO, M. T. *et al.*, "A framework to analyse gender bias in epidemiological research", *Journal of Epidemiology and Community Health*, vol. 61, No. 2, 2007, p. 46.
[559] CHAR, D. S., SHAH, N. H. & MAGNUS, D., "Implementing machine learning in health care – addressing ethical challenges", *The New England Journal of Medicine*, vol. 378, No. 11, 2018, pp. 981-983.
[560] EAGLIN, J. M., "Constructing recidivism risk", *Emory Law Journal*, vol. 67, No. 1, 2017, p. 61.
[561] KEHL, D., GUO, P. & KESSLER, S., "Algorithms in the criminal justice system: assessing the use of risk assessments in sentencing", *Responsive Communities Initiative, Berkman Klein Center for Internet & Society, Harvard Law School*, 2017, p. 2.
[562] *Idem*, p. 24.
[563] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, p. 24.
[564] *Idem*, p. 27.

predicted recidivism in white defendants.[565] However, the firm that designed the algorithm argued that it predicted recidivism risk accurately.

Without entering into the actual discussion regarding whether the algorithm was in fact accurate or not, what is important to highlight is that, considering the historical and existing racial bias in the US law enforcement and justice system,[566] this apparent accuracy in the recidivism score is likely to simply be reproducing past biases and being validated by on-going biases.

Consequently, the way in which the algorithm is built plays a huge role in the use of recidivism algorithms. These algorithms are partly based on the criminal history records of offenders.[567] This is important because, since black people have a higher probability of being arrested, if the programmers choose to, for instance, use "number of prior arrests" as a relevant feature, the risk score of black individuals will be, on average, higher than the average score provided to white people.[568] The way in which algorithms are built is very important, especially when the algorithm is fed biased data.

Furthermore, elements other than criminal history records considered by recidivism algorithms are drawn from tests answered by defendants.[569] The questions in the tests do not ever ask about a person's race.[570] However, within surveys it is possible to find questions such as: "'Was one of your parents ever sent to jail or prison?' 'How many of your friends/acquaintances are taking drugs illegally?' and 'How often did you get in fights while at school'"[571] or "how many prior convictions have you had."[572]

The framework used to develop these questions is clearly skewed. While no specific questions on race or economic status are asked it is clear that the answers provided by a convicted criminal brought up by a well-off family in an affluent neighbourhood will tend to be on the lower-risk side of the scale than those given by someone from a poorer

[565] ANGWIN, J. *et al.*, "Machine bias…", *cit*., 2016.
[566] GROSS, S. R., POSSLEY, M. & STEPHENS, K., "Race and wrongful convictions in the United States", National Registry of Exonerations, 7th May 2017; BERTRAND, M., MULLAINATHAN, S. & ABRAMS, D., "Discrimination in the judicial system", Innovations for Poverty Action, 2001.
[567] DRESSEL, J. & FARID, H., "The accuracy, fairness, and limits of predicting recidivism", *cit.*, 2018, p. 1.
[568] EAGLIN, J. M., "Constructing recidivism risk", *cit.*, 2017, p. 97.
[569] ANGWIN, J. *et al.*, "Machine bias…", *cit.*, 2016.
[570] *Ibidem.*
[571] *Ibid.*
[572] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, p. 25.

background.[573] As it turns out, the latter are generally people from racial minorities who, amongst other things and, as multiple studies show, are more likely to be stopped by the police.[574] Furthermore, given the fact that the risk score is drawn from such a large pool of questions, even when a person with a privileged background has prior convictions, the answers to the rest of the questions will, in many cases, lower their risk score.[575]

In addition, it is important to highlight that one of the recidivism risk tools used by several US prison systems, directly asks questions regarding offenders' economic status such as "How often do you have barely enough money to get by?"[576] The example of recidivism algorithms therefore also offers a deeply worrying picture regarding the general acceptance that appears to exist in using individuals' economic welfare as an element which is directly linked to their risk of recidivism.

3.1.3. Example 3: MS Tay bot

An example of on-going social prejudices, biases and discriminatory attitudes influencing algorithms took place with the launch of MS Tay by Windows in 2016. MS Tay was a machine learning bot that was supposed to be able to interact online with people and learn from said interaction in order to achieve more human responses and thus be able to have conversations. The bot was launched on twitter and published more than 93 thousand tweets before it was disabled twenty-four hours later.[577] The reason behind it being disabled is that after a short period of time, it started to tweet racist, homophobe and misogynistic statements.[578]

There are two elements that must be considered with regard to this example. Firstly, the algorithm was content-neutral and was therefore vulnerable to developing any kind of attitude including socially undesirable behaviours.[579] Secondly, many of the users who interacted with the bot, purposefully exhibited these undesirable attitudes when interacting with MS Tay with the objective of teaching it to adopt them and thus publish discriminatory

[573] *Ibidem.*
[574] NEW YORK CIVIL LIBERTIES UNION, "Stop-and-frisk 2011", 2012.
[575] ANGWIN, J. *et al.*, "Machine bias…", *cit.*, 2016.
[576] NORTHPOINTE, "Risk assessment". Available on 27th March 2019 at: https://www.documentcloud.org/
[577] NEFF, G. & NAGY, P., "Talking to bots: symbiotic agency and the case of Tay", *International Journal of Communication*, vol. 10, 2016, p. 4916.
[578] MONASTERIO ASTOBIZA, A., "Ética algorítmica…", *cit.*, 2017, p. 207.
[579] NEFF, G. & NAGY, P., "Talking to bots…", *cit.*, 2016, pp. 4921-4922.

statements.[580] However, while the users interacting with MS Tay published discriminatory tweets on purpose in a seemingly joking manner, these attitudes are constantly present all over the Internet, meaning that many people –from whose behaviours machine learning algorithms will sometimes learn- still hold these socially undesirable values.[581]

It is therefore essential to design algorithms not as content neutral machines, but considering the social context in which they are developed and deployed. Ultimately, it is necessary for the notion of "equality by design" to act as a mandate in the development of automated systems.[582]

3.2. THE ROLE OF DATA SCIENTISTS: STRUCTURING THE TECH INDUSTRY THROUGH NARRATIVES OF OPPRESSION

As the previous pages convey, there are multiple stages throughout the model creation process that allow for some sort of bias that could later lead to the discrimination of traditionally oppressed groups to be introduced. As well as the stages described above, data scientists can also make many other choices such as eliminating outliers in the data,[583] the degree to which they tune the model, which can allow them to change some of its internal operations while it is being trained,[584] and the types of assessment methods used in order to ensure that the model being trained is as accurate as possible.[585]

Western societies have formally overcome most of the narratives of domination that were previously discussed. Most legal systems now recognise the same rights to those individuals belonging to traditionally oppressed groups as to those who have always been considered full citizens. The existence of formal equality thus leads to the generalised idea that these forms of oppression are in the past.[586] This idea is heavily reinforced by the fact that liberalism places the individual, as an autonomous being, at the centre of all social relations, which helps to ignore the different forms of historical group identity oppression that still underlie

[580] *Ibidem*.
[581] POLAND, B., *Haters: harassment, abuse and violence online*, Lincoln, Potomac Books, 2016.
[582] XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *cit*., 2020, p. 179.
[583] LEHR, D. & OHM, P., "Playing with the data…", *cit.*, 2017, p. 684.
[584] *Idem*, p. 696.
[585] *Idem*, p. 698.
[586] BILGE, S., "Saving intersectionality from feminist intersectionality studies", *Du Bois Review*, vol.10, No. 2, 2013, p. 407.

social structures and that still result in the discrimination of members of disadvantaged groups.[587]

The fact that dominant narratives have shaped the way in which social power structures have been built for centuries means that individuals who belong to the identity groups that have traditionally been in power have managed to accumulate more socioeconomic and political resources. Therefore, traditionally oppressed groups, when recognised full and equal rights, enter a series of power structures that are designed to accommodate the needs of the stereotypical individual shaped by liberal (and now neoliberal) theory and that have not been adapted to their needs, thereby hampering their chances at full inclusion. Furthermore, this political disadvantage is combined with a situation of socioeconomic disadvantage that, in a capitalist society, does nothing but hinder the possibilities that individuals whose identity crosses along one or several of the axes of oppression have in shaping social norms in order to accommodate them as full citizens with real equal opportunities.

That these forms of oppression still take place, although in a much more subtle way, is partly proven by the overrepresentation of autonomous white cisgendered heterosexual males in positions of power. This is specially relevant regarding the topic here addressed given the increasing role that the tech industry has in shaping the way society works. As algorithms and computer technology in general occupy an increasingly central role in Western societies, those who work in the technological sector will have more power in shaping social norms. Thus, the fact that specially protected groups are largely underrepresented in said sector means that their specific needs will be more likely ignored and that traditional narratives of oppression will more easily find the way into the computer systems and technological tools built.[588]

The pervasiveness of dominant narratives is reflected in the fact that when a certain economic sector acquires more importance in society it is rapidly shaped according to dominant norms. This is exactly what happened in the technological sector. The first coders who started to work in the 20th Century were women[589] whose tasks were considered unimportant and were assimilated to administrative or secretary work also generally developed by women, thus

[587] *Ibidem.*
[588] NOBLE, S. U., *Algorithms of Oppression…*, *cit*., 2018, p. 80.
[589] ESMENGUER, N. L., ASPRAY, W. JR. & MISA, T. J., *Computer Boys Take Over: Computers, Programmers, and the Politics of Technical Expertise*, Cambridge (Massachusetts), MIT Press, 2010, pp. 14-15.

being undervalued.[590] However, as the field of computer science started to develop and gain importance, the number of women in this industry began to decrease.

The plunge in female enrolment in computer science degrees is especially illustrative. For example, in Spain, there has been a general downward trend since the 1985-1986 academic year, with the exception of a slight increase in the 1998-1999 academic year. While in 1985-1986 there was a 30% of female enrolment, said proportion was of 12% in the 2016-2017 academic year.[591] The reduction of women enrolled in computer science has been generalised in Western countries[592] and seems to be the result of the way in which gender norms regarding this sector have been constructed.[593] The decrease in female enrolment took place at around the same time that computers started to be commercialized as domestic products whose target audience was specifically male.[594] Additionally, more or less at around the same time aptitude and personality questionnaires started to be used when hiring programmers. These tests privileged male-associated values and the typically masculine educational experience, thus hindering the access of women to that profession.[595] While, the causal relationship between these events has not been statistically proven, it is at least worth mentioning that this evolution in the proportion of women in IT took place and the circumstances under which it did.

Members of disadvantaged groups are still largely underrepresented in the main tech industry sector companies, especially in technical roles.[596] The same narrative that contends that formal equality suffices in order to include members of traditionally oppressed groups as full

[590] *Idem*, p. 30.
[591] MELERO GUERVÓS, J. J. & MERELO MOLINA, C., "Evolución de la matrícula femenina en el grado de informática en universidades públicas españolas", 2017. Available on 26th April 2019 at: https://www.researchgate.net/
[592] ESMENGUER, N. L., ASPRAY, W. JR. & MISA, T. J., *Computer Boys Take Over…*, *cit.*, 2010, p. 237. The decrease in female presence in computer science degrees is not however a characteristic of less developed countries. See, for example, SCHINZEL, B., "Cultural differences of female enrolment in tertiary education in Computer Science", in BRUNNSTEIN, K. & BERLEUR, J., (eds.), *Human Choice and Computers: Issues of Choice and Quality of Life in the Information Society*, Berlin, Springer, 2002, pp. 283-292.
[593] FISHER, A., MARGOLIS, J. & MILLER, F., "Undergraduate women in computer science: experience, motivation and culture", *SIGCSE '97 Proceedings of the twenty-eighth SIGCSE technical symposium on Computer science education*, 1997, pp. 106-110.
[594] HENN, S., "When women stopped coding", *Planet Money,* 21st October 2014. Available on 26th April 2019 at: https://www.npr.org/
[595] ESMENGER, N., "Making programming masculine", in MISA, T. J., (ed.) *Gender codes: Why Women are Leaving Computing*, Hoboken (New Jersey), John Wiley & Sons, 2010, pp. 115-142.
[596] GOOGLE, "Google diversity annual report 2018", 2018, p. 18; MCINTYRE, L., "Diversity and inclusion update: The journey continues", *Microsoft*, November 14th 2018; WILLIAMS, M., "Facebook 2018 diversity report: reflecting on our journey", *Facebook Newsroom*, 12th July 2018. Available on 27th April 2019 at: https://newsroom.fb.com/; APPLE, "Inclusion and diversity", 2017. Available on April 27th 2019 at: https://www.apple.com/diversity/

members of society accordingly argues that a diverse representation in power structures is not necessary either to achieve said objectives. However, the fact that members of traditionally oppressed groups are constantly being discriminated against in said sector,[597] that the narratives of oppression are openly supported by members of the tech industry[598] and the ongoing reproduction of hierarchies of power through the tools developed by said sector[599] proves how formal equality does nothing but reinforce existing power hierarchies.[600] In fact, even when women and individuals who pertain to non-white race groups do enter the tech industry they endure worse working conditions than their white male counterparts and, in many cases, end up leaving said companies due to the discriminatory treatment (including harassment) they are subjected to.[601]

An element that must therefore be necessarily acknowledged is the human factor in creating discriminatory algorithms and models and how the lack of diversity in the computer science workforce hinders a plural perspective from being introduced in the development of machine learning models.[602] It is hard to believe that a gender and race aware perspective[603] will be held by every single member of a tech team such as Google's in which the distribution by gender is 21.4% female and 78.6% male and the distribution by race is 50,7% white, 41.1% Asian, 1.5% black, 2.8% Hispanic, 0.2% Native American and 3,6% of people with two or more races.[604]

For this reason, the role and responsibility data scientists have when avoiding (or creating) discriminatory results from being produced by automated decision-making systems must not be downplayed. It is important to also keep in mind that not only data scientists must be held responsible but also the organisations they work for or anyone who might have some decision in the model's design and deployment, especially considering that some of the stages are

[597] NOBLE, S. U., *Algorithms of Oppression…*, *cit*., 2018, p. 80.
[598] *Idem*, p. 2: "…in the midst of a federal investigation of Google's alleged persistent wage gap, where women are systematically paid less than men in the company's workforce, an 'antidiversity' manifesto authored by James Damore went viral in August 2017, supported by many Google employees, arguing that women are psychologically inferior and incapable of being as good at software engineering as men, among other patently false and sexist assertions."
[599] NOBLE, S. U., *Algorithms of Oppression…*, *cit*., 2018, p. 80.
[600] *Idem*, p. 48: "The internet is in fact organized to the benefit of powerful elites".
[601] KAPOR CENTER, "Tech workforce". Available on 22nd January 2020 at: https://leakytechpipeline.com/
[602] CRAWFORD, K., "Think again…", *cit.*, 2013; US EXECUTIVE OFFICE OF THE PRESIDENT, "Artificial intelligence, automation and the economy", *cit.*, 2016, p. 29; HOUSE OF COMMONS SCIENCE AND TECHNOLOGY COMMITTEE, "Algorithms in decision-making", 2018, p. 22.
[603] While there are many other attributes that are protected by antidiscrimination legislation such as sexual orientation, religious, age…etc., only race and gender are mentioned at this point given that tech company diversity reports generally only include or mainly focus on said attributes.
[604] GOOGLE, "Google diversity annual report 2018", 2018, p. 18.

sometimes outsourced.[605] Such responsibility should be translated to a constant awareness at every stage of the model creating process of the different elements that may cause the final model to yield discriminatory results, thereby reinforcing pre-existing patterns of oppression towards certain groups of people.

The following section contains a series of examples that show how big tech algorithms reinforce structures of disadvantage and subordination.

3.3. DISCRIMINATION DISCOURSES EMBEDDED IN SOCIETY AND THE TECH SECTOR

3.3.1. The sexist and racist nature of Google search

3.3.1.1. The UN's "autocomplete truth" campaign

Over the past few years, a variety of researchers and institutions have brought forward the presence of sexism and racism in Google searches. Perhaps the most prominently known example is the United Nations' "Autocomplete truth" campaign in which Google search autocomplete suggestions showed sexist results when the search began with phrases such as "women shouldn't…", "women should…" or "women cannot".[606] For example, for searches beginning with the phrase "women should…" the autocomplete suggestions shown in the campaign were: "…stay at home", "…be slaves", "…be in the kitchen", "…not speak in church".[607]

The autocomplete examples above conveyed reinforce gender stereotypes that have been developed throughout the course of history which associate women to caregiver roles and with submissive attitudes[608] and which justify silencing women in order to subject them to an androcentric power structure.[609] The campaign placed emphasis in showing the persistence of sexism in society seeing as it focused on how these autocomplete options were the consequence of the searches carried out by Google users.[610] However, by focusing exclusively on the role of users,

[605] LYTVYNOVA, K., "Machine learning project structure: stages, roles, and tools". Available on 7th April 2019 at: https://datafloq.com/
[606] UN WOMEN, "UN Women ad series reveals widespread sexism", 21st October 2013. Available on 10th April 2019 at: http://www.unwomen.org/
[607] *Ibidem*.
[608] FERNÁNDEZ RUÍZ-GÁLVEZ, E., *Igualdad y Derechos Humanos*, *cit.*, 2003, pp.160-161.
[609] FREIXAS, L., "La mujer callada de todos es alabada", speech delivered at the *IV Conference Feminario*, Valencia, 2019. Available on April 10th 2019 at: https://www.youtube.com/
[610] NOBLE, S. U., *Algorithms of Oppression…*, *cit*., 2018, pp. 15-16.

it failed to draw attention to the way the algorithms in Google search are structured in order to allow for these results to show up.[611]

3.3.1.2. Searching for black-sounding names on Google

Another example which underscores the biases generated by search results was put forward by SWEENEY when, in 2013, she published a paper which reflected how introducing black-sounding names in Google was more likely to produce advertisements for checking arrest records in the results than searches made for white-sounding names.[612] As it turned out, while advertisers of said arrest record check products had not ordered and were not looking for their advertisements to be triggered mainly by the search of black-sounding names, the Google search engine ranking system still decided that these results were more relevant in these cases.[613]

Although it is not clear how this decision is made, it is suspected that one of the main elements behind the way in which the results are ranked is the number of times a given result for a certain search has been clicked on by users.[614] Thus in this case, the algorithm would be in fact learning from the racial biases held by Google users. However, considering the Google's search engine algorithm considers more than two hundred different elements when deciding how to rank websites,[615] it is very hard to actually know what the logic behind said results was.

3.3.1.3. Reinforcing gender and racial stereotypes

The risk that results with a high degree of sexually related content will show up when carrying out certain searches on the Internet, even when the user has not previously searched for that type of content, is something that society as a whole has gotten used to and even accepted.[616] As it turns out, these types of results have been especially dominant in searches related to women, thereby exacerbating a social structure that heavily sexualises the female body, reducing women and girls to mere objects.[617] Furthermore, the perpetuation of the traditional gender oppression narrative

[611] *Ibidem*.
[612] SWEENEY, L., "Discrimination in online ad delivery", *Communications of the ACM*, vol. 56, No. 5, 2013, p. 47.
[613] *Idem*, p. 52.
[614] *Ibidem*.
[615] SEOMARK, "How does Google rank websites?", *SEOmark*, 20th September 2019. Available on 11th April 2019 at: https://www.seomark.co.uk/
[616] NOBLE, S. F., *Algorithms of Oppression…*, *cit*., 2018, p. 18.
[617] *Idem*, p. 71.

through Internet searches worsens in cases of identities subjected to intersectional discrimination.[618]

While these results may be partly attributed to a racist and misogynistic social structure, the lack of diversity training these algorithms undergo, largely due to the homogeneity of the tech workforce, must also be highlighted. A clear example of this is the fact that Google Photos has mistaken black people for gorillas when labelling pictures.[619]

Due to the amount of attention drawn to the discriminatory results produced by Google search, specially with regard to racial and gender bias, over the past few years Google has modified its search algorithms so that they produce less racist and sexist results.[620] These modifications in the algorithms tend however to happen only when these situations have become viral and have thus generated a high degree of social outrage.[621] Furthermore in the case, of Google mislabelling pictures of Black people as 'gorillas', the solution provided was simply to eliminate all possibility that any picture would be labelled as "gorilla", "chimpanzee" or "monkey", even pictures of those very animals.[622]

The way disadvantaged groups are portrayed on the Internet (and mainly Google) and the role this portrayal has in reinforcing the existing narrative and discriminatory practices towards said individuals must not be downplayed. Firstly, because the Internet is currently one of the main, if not the primary, source of information for many people throughout the whole world. Secondly, because Google is (by far) the most used search engine and effectively acts as a monopoly.[623] Finally, and most importantly, because the way in which people and groups are represented in the media have been shown to heavily influence the social perception held towards those groups.[624] Moreover, the specific way in which search rankings are set up has been found to manipulate

[618] *Idem*, pp. 93-95. NOBLE, specifically focuses on the narrative surrounding black women and girls.
[619] DOUGHERTY, C., "Google photos mistakenly labels Black people 'gorillas'", *The New York Times BITS Blog*, 1st July 2015. Available on 11th April 2019 at: https://bits.blogs.nytimes.com/
[620] NOBLE, S. F., *Algorithms of Oppression…*, *cit*., p. 10.
[621] *Idem*, p. 80.
[622] SIMONITE, T., "When it comes to gorillas, Google photos remains blind", *Wired*, January 11th 2018. Available on 11th April 2019 at: https://www.wired.com/
[623] CAPALA, M., "Global search engine market share for 2018 in the top 15 GDP nations", 27th August 2018, *Alphametic*. Available on 10th April 2019 at: https://alphametic.com/
[624] PUNYANUT-CARTER, N. M., "The perceived realism of African American portrayals on television", *Howard Journal of Communications*, vol. 19, No. 3, 2008, p. 251; DILL, K. E. & THILL, K. P., "Video game characters and the socialization of gender roles: young people's perceptions mirror sexist media depictions", *Sex Roles: A Journal of Research*, vol. 57, No. 11-12, 2007, pp. 851-864.

individuals.[625] Consequently, the relevance of these cases, in which the perpetuation of discriminatory attitudes towards specially protected groups is conveyed, must also be considered when addressing the discriminatory results generated by algorithmic decision-making.

3.3.2. Profiling algorithms to exclude disadvantaged groups in targeted advertising

One of the ethical dilemmas that arise from the growing use of personal data by private big tech firms is the fact that individuals who are users of online platforms are able to access their products for free because "users" are the product. Companies such as Facebook or Google obtain most of their revenue from advertising.[626] In addition to the general ethical problems of treating humans as products that some scholars and activists have pointed out, targeted ads can also help perpetuate situations of disadvantage.

The advantage of posting an advertisement on one of these platforms is that they can use algorithms to select those platform users that could be interested in the advertisement. Targeted advertising entails showing certain advertisements only to those users who have a set of given features selected by the advertiser.[627] Advertising engines, such as Google Ads, do so by analysing data obtained from different sources such as web browsing histories, social network profiles and Google Ad Settings amongst others.[628] As with many of the examples included in this work, targeted advertising does not necessarily produce negative or discriminatory outcomes. In fact, it can be very useful seeing as it can help consumers find products they are really interested in, not having to search amongst the vast amount of information available on the Internet. However, there have been cases in which individuals belonging to certain groups have been excluded from certain advertisements.

DATTA *et al.*[629] carried out an experiment in which they controlled for the differences in the types of advertisements shown to female and male job-seekers. They did so by having two groups of people select either "male" or "female" on the gender option in the Google Ads settings page and

[625] EPSTEIN, R. & ROBERTSON, R. E., "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections", *PNAS*, vol. 112, No. 33, 2015, pp. E4512-E4521.
[626] ALPHABET, "Alphabet Announces Second Quarter 2020 Results", 2020. Available on 12th September 2020 at: https://abc.xyz/; FACEBOOK, "Facebook Q2 2020 results", 2020.
[627] SPEICHER, T. *et al*., "Potential for discrimination in online targeted advertising", *Proceedings of Machine Learning Research*, 81, 2018, p. 1.
[628] GOOGLE AD SETTINGS. Available on 11th April 2019 at: https://adssettings.google.com/: "Ads are based on personal info you've added to your Google Account, data from advertisers that partner with Google, and Google's estimation of your interests. Choose any factor to learn more or update your preferences."
[629] DATTA, A., TSCHANTZ, M. C. & DATTA, A., "Automated experiments on ad privacy settings: a tale of opacity, choice and discrimination", *Proceedings on Privacy Enhancing Technologies 2015*, 2015, p. 93.

then browse employment-related website. The result was that males were displayed advertisements for a coaching firm which promised large salaries at a much higher frequency than females. In this case, it was not clear whether the discriminatory outcomes resulted from the firm posting the advertisements having specified that it preferred male to female candidates or from the Google Ads' algorithm deciding that women would not be interested in said advertisements.

Another case in which disadvantaged groups were excluded from accessing goods or services took place in Facebook's online marketplace. The social platform used profiling in order to offer advertisers the possibility of excluding ethnic minorities from viewing certain advertisements by including an option labelled "ethnic affinities" in its advertisement settings.[630] By selecting all or several of the items within this option, advertisers could exclude people with the chosen ethnic affinities from being shown the ad.[631] In the research carried out by ANGWIN and PARRIS JR., which focused on housing advertisements, the ethnic affinity categories that could be excluded were: African American, Asian American and Hispanic,[632] all non-white communities.

Facebook explained that "ethnic affinities" did not mean "ethnic group" but the interest that an individual had shown in content specifically related to a certain ethnic group.[633] However, considering this option was introduced under the "demographics" settings it is quite clear that the objective was to allow advertisers to exclude people belonging to certain ethnicities. Furthermore, what this example comes to show is that algorithmic profiling is being actively and consciously carried out by digital platforms and, although the final decision concerning the discrimination of certain groups is not made by algorithms in this case, the use of automated tools when drawing up profiles of people which include specially protected attributes such as ethnicity, clearly proves how these technologies offer new ways in which to stream discrimination and how, for the time being, equality considerations are not being introduced in digital environments.

Facebook committed to changing its Ad settings so that they would not allow for discriminatory options and, in a public statement,[634] indicated that it had put in place "stronger enforcement tools" that would flag any advertisements regarding housing, employment or credit opportunities

[630] ANGWIN, J. & PARRIS JR., T., "Facebook lets advertisers exclude users by race", *Propublica*, 28th October 2016; MONASTERIO ASTOBIZA, A., "Ética algorítmica…", *cit.*, 2017, p. 202; SPEICHER, T. *et al.*, "Potential for discrimination in online targeted advertising", *cit.*, 2018, p. 2.
[631] ANGWIN, J. & PARRIS JR., T., "Facebook lets advertisers exclude users by race", *cit.*, 2016.
[632] *Ibidem.*
[633] ANGWIN, J. & PARRIS JR., T., "Facebook lets advertisers exclude users by race", *cit.*, 2016; SPEICHER, T. *et al.*, "Potential for discrimination in online targeted advertising", *cit.*, 2018, p. 2.
[634] FACEBOOK, "Improving enforcement and promoting diversity: updates to ads policies and tools", *Facebook Newsroom*, 8th February 2017.

that discriminated on the basis of "multicultural affinity"[635] seeing as it was a clear proxy for race. However, a second piece of research carried out proved that, Facebook was still allowing housing advertisements not be shown to people with certain "multicultural affinities".[636] Furthermore, it was proven that Facebook also allowed excluding people interested in "gay men", "soccer moms" and "sign language".[637]

Facebook finally introduced the necessary changes regarding targeted advertising on housing in March 2019 so that it would not be possible to discriminate on the basis of any protected attribute in advertisements related to employment, housing or credit.[638]

3.4. POLICY CHOICES

In many cases, algorithms are not themselves discriminatory but it is the decisions made regarding how they are used that lead to discriminatory outcomes. For example, since 2018, the Austrian public Employment Services has been using an algorithm on a trial basis in order to classify unemployed individuals according to their probability of finding a new job. The final introduction of this automated system, called PAMAS, was approved to take place during 2020 and its objective is to carry out public resource assignation in the most efficient way possible.[639]

The algorithm analyses different characteristics of unemployed individuals and provides each person with a score. Once the scoring process has taken place, the algorithm classifies unemployed people in three different categories according to the probabilities they have of finding a new job: high, medium or low probability.[640]

The first controversy surrounding this automated system arose when the firm hired to create the model published the source code and underlying logic. In said document it was possible to see how points were taken away if certain characteristics such as being a woman or a non-

[635] After ANGWIN & PARRIS' research was published, Facebook changed the "ethic affinity" setting to "multicultural affinity" and included it under "Behaviors" instead of "Demographics. ANGWIN, J., TOBIN, A. & VARNER, M., "Facebook (still) letting housing advertisers exclude users by race", *Propublica*, 21st November 2017.
[636] *Ibidem.*
[637] *Ibid.*
[638] GILLUM, J. & TOBIN, A., "Facebook won't let employers, landlords or lenders discriminate in ads anymore", *Propublica*, 19th March 2019.
[639] SZIGETVARI, A., "Arbeitsmarktservice gibt grünes Licht für Algorithmus", *Der Standard*, 17th September 2019. Available on 23rd January 2020 at: https://www.derstandard.at/
[640] OECD, "Profiling tools for early identification of jobseekers who need extra support", December 2018, p. 3.

EU member state national concurred in an individual. It thus became obvious that, in general, individuals' scores would be lowered as they qualified into more groups at risk of being discriminated. Hence, vulnerable individuals with a greater risk of social exclusion would be classified into the "low probability of finding a job" category.

While this system is objective in that it simply reflects the discriminatory practices that exist in the labour market, it has been heavily criticised by certain Austrian social organisations and sectors. These criticisms are not without reason given the fact that just by classifying members of vulnerable groups in the lower category, the system contributes to the stigmatisation of these groups and their members.[641] The Austrian public Employment Services defended the need for the system to operate in this manner in order to ensure that public resources are distributed in the best way possible, providing more adequate help to those individuals who might find more difficulties in accessing the job market with more appropriate help.[642]

However, while the Austrian public Employment Services justified this form of classification under the pretext of offering better help to individuals who had more difficulties in finding a new employment, they decided to prioritise efficiency in the allocation of public resources over any other objective. Hence, after reaching the conclusion that the most efficient resource allocation would be providing more resources to individuals with a medium probability of finding a new job, the Austrian public Employment Services decided to considerably reduce the amount of resources and other aid provided to unemployed individuals whose chances of reentering the job market are lowest,[643] therefore perpetuating the situations of social exclusion suffered by certain groups and individuals, as well as helping to reinforce the construction of social structures and institutions through narratives of subordination of historically oppressed groups.

Additionally, welfare service allocation is increasingly being automatised.[644] If automated welfare programmes are not set up in order to ensure that the fundamental rights to privacy

[641] ZARSKY, T., "Understanding discrimination in the scored society", *cit*., 2014. The discussion regarding whether this form of classification should be tolerated and accepted or whether it should be banned in order to prevent the risk that it acts as a trigger to reinforce the discrimination undergone by vulnerable groups in the employment market will be approached to a greater extent in the final part of the thesis.
[642] PLANET LABOR, "Austria: an algorithm that evaluates the unemployed (briefly)", 24th October 2018. Available on 23rd January 2020 at: https://www.planetlabor.com/
[643] OECD, "Profiling tools for early identification of jobseekers who need extra support", December 2018; SZIGETVARI, A., "Arbeitsmarktservice gibt grünes Licht für Algorithmus", *cit.*, 2019.
[644] ALSTON, P., "Digital welfare states and human rights", *cit.*, 2019.

and due process of welfare recipients are being properly respected, these systems can easily worsen the vulnerable situations of certain population groups. This is especially when it comes to aid programmes aimed towards poverty relief. Poor individuals have fewer resources and can therefore, in general, find it more difficult to challenge decisions and confront situations in which they are treated unfairly. These situations can be worsened if decisions are made by machines which hinder even further knowledge on how the decision was made or what the review mechanisms are.

In relation to the automatisation of welfare programmes it is especially relevant to point out the increasing privatisation of welfare fraud detection and prediction systems. This process is driving social security systems away from their aim as safety nets for citizens.[645] Attention is increasingly being placed on controlling the poor, thereby reinforcing negative stereotypes of members of lower socioeconomic strata as untrustworthy "others" that take advantage of state provided services and aid[646] and who have to be controlled and whose rights to privacy and data protection only exist to a lesser degree than for the rest of citizens.[647]

4. ALGORITHMS CAN PERPETUATE SOCIAL STRUCTURES OF DISCRIMINATION

The fact that machines may develop autonomous trains of thought poses obvious risks since, even if it is not clear how, they might draw correlations between the data they are fed that might lead to the discrimination of especially vulnerable groups. The lack of transparency[648] of these systems becomes more problematic in the case of deep learning than of machine learning since the decisions are made based on layers upon layers of largely self-developed knowledge.[649] In fact, as it will be later discussed to much greater length in part II, the lack of transparency that sometimes characterises decision-making models can help to "obscure undesirable behaviour",[650] in other words, it can help to hide discrimination that is willingly

[645] RANCHORDÁS, S. & SCHUURMANS, Y., "Outsourcing the welfare state…", *cit.*, 2020, pp. 5-42.
[646] EUBANKS, V., *Automating Inequality…*, *cit.*, 2017, pp. 14-16.
[647] RANCHORDÁS, S. & SCHUURMANS, Y., "Outsourcing the welfare state…", *cit.*, 2020, p. 40.
[648] This is an issue that will be approached in much more detail when analysing the different proposals for regulating algorithmic-based discrimination in Part II of the thesis.
[649] GOODMAN, B. W., "A step towards accountable algorithms?: Algorithmic discrimination and the European Union general data protection", paper presented at the *29th Conference on Neural Information Processing Systems*, Barcelona, 2016, p. 3. Available on 13th February 2019 at: http://www.mlandthelaw.org/
[650] HOUSE OF COMMONS SCIENCE AND TECHNOLOGY COMMITTEE, "Algorithms in decision-making", 2018, p. 18.

introduced in the algorithms and models by their programmers or any other actor involved in its creation in some way.

Furthermore, one of the main problems with machine learning models which has already been stated is the fact that they may not only learn from past biases but, once the model is working, seeing as it keeps on learning it could also encode existing prejudice and, even if constraints are introduced so that models do not use certain personal characteristics or identities, the amount of data that these systems are fed and the fact that they work by making correlations entails that it is very difficult to actually avoid the inclusion of biases at some point during their creation, training or deployment.

Additionally, the reason why big data has become so relevant both for public and private organisations is not the data itself but the information that results from the correlations and inferences that can be drawn when processing it through algorithms.[651] Consequently, organisations designing automated decision-making algorithms may design said mechanisms in order to obtain as many patterns as possible, sometimes even disregarding the possibility that said patterns could in fact have no significance.

As it has been conveyed, discrimination can be easily introduced –inadvertently or on purpose- in machine learning algorithms and it is in fact quite difficult to point out the underlying logic behind algorithmic discrimination or, if it is the case, when and where in the design and creation process was discrimination encoded in the automated decision-making model. However, this does not mean that the processes behind algorithmic discrimination should be ignored. Increased literacy in the new technologies here analysed is absolutely necessary if legal scholars and regulators want to introduce at least some degree of control in the model creation process and set some firewalls that can prevent to a certain degree the introduction of biases in the tool.

Consequently, a general overview of the ways in which automated decision-making works is vital in order to build up an explanation of the ways in which algorithms may produce discriminatory results and design new regulations aimed towards controlling and preventing algorithmic discrimination but also determine how existing regulations can be applied to these instruments. Based on the theoretical framework set in the previous chapter and the

[651] ILLINGWORTH, A. J., "Big data in I-O psychology: privacy consideration and discriminatory algorithms", *Industrial and Organizational Psychology*, vol. 8, No. 4, 2015, p. 570.

practical explanations and examples of how algorithms discriminate, the following chapter sets out to address the ways in which the European equality and non-discrimination framework can address algorithmic discrimination.

CHAPTER IV. APPLYING THE EU EQUALITY AND ANTI-DISCRIMINATION FRAMEWORK TO ALGORITHMS

This chapter addresses the EU framework for the protection of the rights to non-discrimination and equality and its applicability to different forms of algorithmic discrimination. The first section draws a general overview of the European equality and anti-discrimination framework. The second section focuses on what constitutes direct algorithmic discrimination and how it may be proven. The third section follows a very similar structure to the second section but with regard to indirect algorithmic discrimination. The fourth and final section focuses on the European substantive equality framework and how algorithmic affirmative action can help to redress the harms caused by structural discrimination.

1. THE EU EQUALITY AND ANTI-DISCRIMINATION FRAMEWORK

All western democracies and human rights treaties recognise the general principle to equal treatment as a fundamental right, presented as the very essence of liberal regimes, the objective of which is to ensure all individuals are treated equally and therefore granted the possibility of accessing and exercising their individual rights in equal standing with their fellow citizens.[652] This general principle, which rests on the idea that similar situations must be treated similarly unless differential treatment is justified,[653] generally allows for ample justifications for differential treatment, as long as said difference is rationally explained.[654]

In addition to said general principle, the equality regulatory framework also includes more specific prohibitions on discriminatory actions based on certain specific grounds that reflect those elements upon which the dominant narratives of oppression have traditionally been built upon. As it has already been stated, the sub-categories that represent disadvantaged groups within protected grounds are the ones that the dissertation focuses on.

The framework that addresses the rights to equality and non-discrimination based on protected grounds and, in particular, the equality and non-discrimination of disadvantaged

[652] ALLEN, R. & MASTERS, D., "Artificial Intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making", *ERA Forum*, vol. 2020, No. 4, 2020, p. 591.
[653] GELLERT, R. *et al.,* "A comparative analysis of anti-discrimination and data protection legislations", in CUSTERS, B. *et al*., (eds.), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in large Databases*, Berlin, Springer, 2013, p. 64.
[654] GELLERT, R. *et al.,* "A comparative analysis of anti-discrimination and data protection legislations", *cit.*, 2013, p. 65.

groups, does so from two perspectives.[655] On the one hand, through prohibitions on differential treatment that is considered especially harmful[656] due to the fact that it produces unfair decisions and perpetuates and reinforces existing situations of structural discrimination suffered by the members of traditionally disadvantaged groups.[657] This outlook on discrimination aims towards detecting discriminatory practices and enforcing the aforementioned prohibitions.[658] On the other hand, there are also a series of mechanisms that aim towards preventing and/or eradicating discrimination through the implementation of mechanisms such as affirmative action.[659] The combination of both types of approaches which, as the following paragraphs convey, takes place in EU law, means that the European equality and anti-discrimination framework is, at least in theory, mostly built from an anti-subordination perspective.

1.1. EUROPEAN INSTRUMENTS TO PROTECT EQUALITY AND NON-DISCRIMINATION

The European legal framework for the protection of the rights to equality and non-discrimination is embodied in the equality and non-discrimination clauses contained in the European Convention of Human Rights, the Charter of Fundamental Rights of the EU, the Treaty on the Functioning of the EU (TFEU), the Directives on discrimination and the Council Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law.

Article 14 of the European Convention on Human Rights and article 1 in Protocol 12 to the Convention ban discrimination on the grounds of "sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status". The difference between article 14 of the Convention and article 1 of Protocol 12 to the Convention and the very reason why the latter was passed is that article 1 of Protocol 12 expands the material scope of application of the non-discrimination clause. While article 14 of the Convention can only be invoked in relation to

---

[655] BARRÈRE UNZUETA, M. A., "Problemas del derecho antidiscriminatorio: subordinación versus discriminación y acción positiva versus igualdad de oportunidades", *Cuadernos Electrónicos de Filosofía del Derecho*, No. 9, 2003a, p. 6.

[656] For example, article 14 of the European Convention on Human Rights is titled: "Prohibition of discrimination". Also see, NUNZIATO, D. C., "Gender equality: states as laboratories", *Virginia Law Review*, vol. 80, No. 4, 1994, pp. 946: "The antidiscrimination mediating principle is one of negative restraint that forbids the government from arbitrarily discriminating against classes of individuals".

[657] ROBERTS, J. L., "Protecting privacy to prevent discrimination", *cit.*, 2015, p. 2111.

[658] BARRÈRE UNZUETA, M. A., "Problemas del derecho antidiscriminatorio…", *cit.*, 2003a, p. 6.

[659] *Ibidem*.

other rights and freedoms contained in the Convention,[660] article 1 of Protocol 12 removes said limitation and establishes that the right to non-discrimination can be invoked with regard to "the enjoyment of any other right set forth by law". Although the framework of protection of the rights to equality and non-discrimination applicable to the parties of the Convention is analysed, it is important to keep in mind that the objective is to focus on its impact on EU.

Article 21 of the Charter of Fundamental Rights of the EU also establishes the right to non-discrimination on the grounds set by the Convention and Protocol 12 and adds the following grounds: ethnic origin, disability, age and sexual orientation. In neither case are the grounds a *numerus clausus*, hence other criteria used when profiling and making decisions on individuals could also be included under these especially harmful cases of discrimination.[661]

Race is the specific ground that is more comprehensively covered through EU legislation. The Race Equality Directive[662] prohibits discrimination in employment, occupation, vocational training, several areas of social welfare, including social security and education, and access to goods and services.[663] In addition, some the most harmful forms of racial and xenophobic discrimination, specified in incitation to violence against groups defined by race, colour, religion, descent or national or ethnic origin and condoning, denying or trivialising crimes such as genocide, are covered by the EU's Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law.

The rights to not be discriminated based on gender and to equality between women and men are also provided quite an extensive protection (if compared to other specific grounds). Article 8 of the TFEU establishes a general mandate indicating that "the Union shall aim to eliminate inequalities, and to promote equality, between men and women". The Treaty also

[660] Article 14 of the European Convention on Human Rights states that: "The enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status".

[661] SCHREURS, W. *et al.,* "Cogitas, ergo sum. The role of data protection law and non-discrimination law in group profiling in the private sector" in HILDEBRANDT, M. & GUTWIRTH, S. (eds.), *Profiling the European Citizen*, Berlin, Springer, 2008, pp. 259-260.

[662] Council Directive 2000/43/EC implementing the principle of equal treatment between persons irrespective of racial or ethnic origin.

[663] See art. 3 of the Race Equality Directive. It is also important to consider that the Race Equality Directive "does not cover difference of treatment based on nationality and is without prejudice to provisions and conditions relating to the entry into and residence of third-country nationals and stateless persons on the territory of member states, and to any treatment which arises from the legal status of the third-country nationals and stateless persons concerned".

specifically addresses gender equality in employment in articles 153 and 157. Article 153.1.i establishes that the EU "shall support and complement the activities of the member states [in achieving] equality between men and women with regard to labour market opportunities and treatment at work". In addition, article 157.1 establishes and the principle of equal pay without discrimination based on sex. Finally, it is especially relevant to point out that article 157.4 TFEU establishes the basis for implementing positive action as it states that member states may maintain or adopt "measures providing for specific advantages in order to make it easier for the underrepresented sex to pursue a vocational activity or to prevent or compensate for disadvantages in professional careers".

The protection of equality and non-discrimination on the basis of gender in EU secondary law is materialised, on the one hand, in the Gender Employment Equality Directive,[664] which prohibits gender-based discrimination in employment and occupation. However, it has a more limited scope than the Race Equality Directive, as it does not include certain areas such as education and healthcare. The prohibition of sex discrimination in access to and supply of goods and services is also established in what is commonly known as the Gender Goods and Services Directive.[665] Said Directive explicitly excludes media, advertising and education from its scope of application.[666] Finally, the Self-employment Gender Equality Directive[667] prohibits gender-based discrimination in self-employed activities and extends a series of protections, such as maternity leave, to self-employed workers.

The Employment Equality Directive prohibits instances of discrimination in employment on the grounds of religion and belief, age, disability and sexual orientation.[668] Said Directive only covers employment, occupation and vocational training but does not include social security under its scope of application. It is also relevant to highlight that none of the cited regulatory instruments that conform the EU's equality and anti-discrimination framework include social origin, property or any other element that might point to socioeconomic status

[664] Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast).
[665] Council Directive 2004/113/EC of 13 December 2004 on equal treatment for men and women in the access to and supply of goods and services.
[666] Article 3.3 in the Gender Employment Equality Directive.
[667] Directive 2010/41/EU of the European Parliament and of the Council of 7 July 2010 on the application of the principle of equal treatment between men and women engaged in an activity in a self-employed capacity and repealing Council Directive 86/613/EEC.
[668] Council Directive 2000/78/EC establishing a general framework for equal treatment in employment and occupation.

as one of the specially protected grounds regarding which discrimination is more intensely forbidden.

Additionally, there are some elements that can be argued to fall within the special categories of discrimination contained in the Convention and Charter but that do not fall within the scope of the EU Equality Directives. For example, article 3.2 of the Employment Equality Directive and the Racial Equality Directive clearly state that said regulatory instruments do not cover differences in treatment based on nationality. Nevertheless, and given the fact that discrimination on the basis of nationality can be an expression of racial/ethnic discrimination, many EU member states do include nationality as a forbidden ground for differential treatment.[669] Discrimination on the basis of age or disability in the armed forces is also excluded from the scope of application of the Employment Equality Directive.

Finally, it is important to highlight that both the provisions protecting the rights to equality and non-discrimination in the TFEU and article 21 of the Charter of Fundamental Rights of the EU have direct horizontal effect and therefore confer "individuals a right which they may rely on as such in disputes between them in a field covered by EU law".[670]

1.2. THE INDIVIDUALISTIC APPROACH TO DISCRIMINATION

All the aforementioned Directives do not just focus on establishing prohibitions to discrimination, but also set out a system aimed towards the promotion of equality through gender mainstreaming, equality bodies and positive action. Therefore, as it was indicated, the EU regulatory framework draws from anti-subordination perspectives and, in theory, aims to dismantle the structural system of oppression that disadvantages the members of certain groups. However, the EU equality and anti-discrimination system is largely inspired by the US model[671] and, partly due to this, but also generally in line with the legal tradition that has placed the individual as the nucleus, subject of rights and therefore the unit of analysis upon which fundamental rights are structured and made effective, European institutions and, particularly, the Court of Justice of the EU (CJEU), have failed to effectively combine the

[669] CHOPIN, I. & GERMAIN, C., "A comparative analysis of non-discrimination law in Europe 2019", Brussels, European Commission, 2019, pp. 75-77.
[670] CJEU Judgment, 17th April 2018, C-414/16, Vera Egenberger v. Evangelisches Werk für Diakonie und Entwicklung e.V, paragraphs 76-79.
[671] DE BURCA, G., "The trajectories of European and American antidiscrimination law", *The American Journal of Comparative Law*, vol. 60, No. 1, 2012, pp. 4-5; HEPPLE, B., "Equality at work", in HEPPLE, B. & VENEZIANI, B., *The Transformation of Labour Law in Europe: A Comparative Study of 15 Countries 1945-2004*, Portland, Hart Publishing, 2009, p. 161.

group-dimension of discrimination with the notion of formal and individual-based equality that underlies the recognition of the rights to equality and non-discrimination.[672]

Hence, the EU has built and mainly applies its equality and non-discrimination framework following from the idea that treating individuals differently on the basis of one of the protected categories is wrong and, in principle, unjustified but, in many cases, fails to acknowledge that there is more wrong in discriminating against a woman than a man or against a black person than a white person due to the historical and systemic disadvantage that subordiscriminated groups endure. Hence, the implementation of the equality and anti-discrimination model mostly based on notions of formal equality, has led the CJEU to be reluctant to accept different forms of positive action and to effectively incorporate the structural discrimination perspective that is needed in order to properly evaluate cases of discrimination.

Another element to consider, which will be approached in the next chapter, is the fact that, while the US model does incorporate class actions which, to a certain extent help to include the group element of discriminatory structures into the system of judicial remedies, class actions are not generally implemented in the EU and, even when recognised, have seldom been exercised in discrimination cases.[673] Hence, although the EU equality and non-discrimination framework has developed into a more comprehensive system than the one it is inspired on and apparently aims to help eliminate existing systems of structural discrimination, its heavy reliance on conceptualising discrimination as an intersubjective conflict leads to a very weak implementation of the anti-subordination perspective. While we will return to this particular criticism in the next chapter, it is important to keep it in mind throughout the analysis carried out throughout the rest of the chapter.

1.3. THE FOCUS ON EU SECONDARY LAW

This chapter mainly focuses on the EU secondary law framework, that is, the EU Equality Directives, due to the fact that said Directives specifically protect disadvantaged and historically oppressed groups and establish a more comprehensive framework on the way in

[672] AÑÓN ROIG, M. J., "Principio antidiscriminatorio y determinación de la desventaja", *cit.*, 2013b, p. 135; RUBIO, A., "Las políticas de igualdad: de la igualdad formal al mainstreaming", 2003, p. 17. Available on 15th March 2020 at: http://pmayobre.webs.uvigo.es/

[673] CHOPIN, I. & GERMAIN, C., "A comparative analysis of non-discrimination law in Europe 2019", *cit.*, 2019, pp. 92-94.

which the principle of equal treatment should be enacted than the provisions on equality and non-discrimination contained in the Charter of Fundamental Rights of the EU and the TFEU. The EU Equality Directives lay down the framework for combating discrimination in the EU and thus, the backdrop against which the equality and non-discrimination framework should be implemented in all EU member states. In this sense, the legal instruments enacted in the member states that cover the remit of the Equality Directives and which must respect and be interpreted in light of these regulatory instruments.[674]

In any case, the analysis must be understood to be carried out within the framework set by the more general provisions on equality and non-discrimination contained in the Charter and the TFEU. Additionally, ECHR case law is also examined.

Following from the European construction of discrimination through the identification of cases as falling within the categories of direct or indirect discrimination and the possibilities that public powers have in establishing positive action mechanisms for the promotion of equality, the following sections review the way in which said concepts are shaped by the EU regulatory framework, the CJEU and the European Court of Human Rights (ECHR) and the extent to which they can be applied to algorithmic discrimination.

2. DIRECT ALGORITHMIC DISCRIMINATION

Cases of direct discrimination take place when a decision, criterion or practice explicitly discriminates on the basis of a protected characteristic. Within this scope, the discrimination against disadvantaged groups, as sub-categories within protected characteristics, is the main focus of this section as it is of the rest of the dissertation.

2.1. GENERAL REQUIREMENTS

There are three main elements that are commonly applied by European and national courts in order to determine whether non-discrimination regulation must apply to a particular case. These elements are almost inseparable from one another in practice and follow from the definitions of direct discrimination provided, amongst other regulatory instruments, by the European anti-discrimination framework. In order for a direct discrimination case to be

[674] CIACCHI COLOMBI, A., "The direct horizontal effect of EU fundamental rights: ECJ 17 April 2018, Case C-414/16, Vera Egenberger v Evangelisches Werk für Diakonie und Entwicklung e.V. and ECJ 11 September 2018, Case C-68/17, IR v JQ", *European Constitutional Law Review*, vol. 15, No. 2, 2019, pp. 294-305.

considered an individual must be (1) treated less favourably than others have been or would be (2) in a similar situation (3) as a consequence of a protected ground having been taken into consideration in the decision made.

2.1.1. Less favourable treatment and a comparator

Thus, the first element that must be tested for, and proven, is the existence of less favourable treatment.[675] In order to prove the first element, courts generally require claimants to provide the second element, a comparator.[676] For example, a facial recognition system such as the one used in many smartphones which has a significantly greater error rate with Black people than it does with white people may be argued to be a case of less favourable treatment on the basis of race in the access to a good or service.[677] The way in which courts interpret the validity of the case offered as a comparator has a tremendous impact with regard to the extent to which the right to equality and non-discrimination is recognised.[678]

If courts require the comparator to be identical to the alleged discriminatory situation[679] or do not admit previous or hypothetical situations as comparators, the extent to which individuals can claim the existence of cases of discrimination will be deeply hindered. For instance, in gender equal pay cases, a very restrictive interpretation of situations that could be considered valid comparators would lead to women only being able to compare their salaries to male colleagues who developed the exact same tasks in the same company. For example, in the *Allonby v. Accrington & Rossendale College* case, a woman was dismissed from her job as a teacher and then rehired by an intermediary company that placed her in her former workplace to perform the exact same duties as before but with a significant reduction in her wages. The applicant provided as a comparative reference the wages received by male teachers with better pay. However, the CJEU concluded that that comparative reference was not valid since the teachers serving as comparators were employed directly by the establishment in which

[675] CHOPIN, I. & GERMAIN, C., "A comparative analysis of non-discrimination law in Europe 2019", *cit.*, 2019, p. 40.
[676] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", Luxembourg, Publications Office of the European Union, 2018, p. 44.
[677] ALLEN, R. & MASTERS, D., "Artificial Intelligence…", *cit.,* 2020, p. 592.
[678] BALAGUER CALLEJÓN, M. L., "Igualdad y discriminación sexual en la jurisprudencia del TC", *Revista de Derecho Político*, No. 33, 1991, p. 114.
[679] *Ibidem.*

they carried out their work, whereas the applicant was employed by the intermediary company.[680]

However, there have also been cases in which the CJEU has been more flexible in considering that a discrimination claim fell under the scope of direct discrimination even when no comparator was provided. For example, in the *Dekker v. VJV* case[681] the plaintiff had been considered to be the only suitable candidate for a job position. However, as she informed the firm during the recruitment process that she was pregnant, she was refused employment because the company's insurance stipulated that it would not pay for the costs of the candidate's maternity leave. The CJEU found that, since pregnancy was directly linked to the sex of the candidate, the decision not to promote her because she was pregnant constituted direct discrimination on the grounds of gender,[682] regardless of whether there were male candidates with whom to compare the decision made with regard to the applicant.[683] The consideration of unfavourable treatment to pregnant women as constitutive of direct discrimination was later introduced in Recital 23 of the Gender Employment Equality Directive.

[680] CJEU Judgment 13th January 2004, C-256/01, Debra Allonby v. Accrington & Rossendale College, Education Lecturing Services, trading as Protocol Professional and Secretary of state for Education and Employment: "In circumstances such as those of the main proceedings, Article 141(1) EC must be interpreted as meaning that a woman whose contract of employment with an undertaking has not been renewed and who is immediately made available to her previous employer through another undertaking to provide the same services is not entitled to rely, vis-à-vis the intermediary undertaking, on the principle of equal pay, using as a basis for comparison the remuneration received for equal work or work of the same value by a man employed by the woman's previous employer."

[681] CJEU Judgment 8th November, 1990, C-177/88, Elisabeth Johanna Pacifica Dekker v. Stichting Vormingscentrum voor Jong Volwassenen (VJV Centrum) Plus.

[682] *Idem*, paragraph 12: "In that regard it should be observed that only women can be refused employment on grounds of pregnancy and such a refusal therefore constitutes direct discrimination on grounds of sex. A refusal of employment on account of the financial consequences of absence due to pregnancy must be regarded as based, essentially, on the fact of pregnancy. Such discrimination cannot be justified on grounds relating to the financial loss which an employer who appointed a pregnant woman would suffer for the duration of her maternity leave."

[683] *Idem*, paragraphs 15-17: "In its second question the Hoge Raad asks whether the fact that there was no male candidate for the job is liable to alter the answer to the first question. The VJV contends that the second question must be answered in the affirmative, because what is involved is not the discriminatory effect of an abstract measure but a concrete decision by an employer not to engage a specific candidate. When an employer chooses from among exclusively female candidates, his choice cannot be attributable to discrimination on grounds of sex, because in such a case the employer is guided by other considerations of a financial or administrative nature. If that reason is to be found in the fact that the person concerned is pregnant, then the decision is directly linked to the sex of the candidate. In those circumstances the absence of male candidates cannot affect the answer to the first question."

2.1.2. Causal link

The third and final element is "the need for a causal link between the less favourable treatment and the protected grounds",[684] which means that the individual who has undergone a less favourable treatment would not have experienced it had she not pertained to one of the specially protected groups. In some cases, courts have interpreted the causal link to also be generated by elements that are inseparable from the protected ground. For example, if a series of benefits are associated to marriage and homosexuals do not have the right to get married, marriage is inseparable from sexual orientation and the measure would be directly discriminating against homosexuals.[685] This particular case is considered direct discrimination because the element considered for the decision "marriage" effectively reflects the protected ground and does not just correlate to it, in which case it would fall under the scope of indirect discrimination.

2.1.3. Victim status

In order to consider that a case of direct discrimination has taken place, the ECHR demands that the individual who claims that her right to equality has been violated is directly affected by the discriminatory action, that is, that she holds "victim status".[686] Conversely, EU law establishes that even if there is no identifiable victim, discriminatory actions can still be judged. This is due to the fact that, while the ECHR solves specific cases, the CJEU interprets and determines the general application of EU law in member states.

In the *Feryn* case, the CJEU ruled that a instance of direct discrimination had taken place when an employer publicly stated that he would not hire immigrants. The Court considered this would dissuade members of said protected group (falling within the race and ethnicity suspect category) from applying for a job with said employer.[687] In the same vein, the CJEU

[684] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", *cit.*, 2018, p. 49.
[685] CJEU Judgment 1st April 2008, C-267/06, Tadao Maruko v. Versorgungsanstalt der deutschen Bühnen, paragraph 72: "If the referring court decides that surviving spouses and surviving life partners are in a comparable situation so far as concerns that survivor's benefit, legislation such as that at issue in the main proceedings must, as a consequence, be considered to constitute direct discrimination on grounds of sexual orientation, within the meaning of Articles 1 and 2(2)(a) of Directive 2000/78."
[686] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", *cit.*, 2018, p. 43.
[687] CJEU Judgment 10th July 2008, C-54/07, Firma Feryn NV v. Centrum voor gelijkheid van kansen en voor racismebestrijding: "The fact that an employer states publicly that it will not recruit employees of a certain ethnic or racial origin constitutes direct discrimination in respect of recruitment within the meaning of Article 2(2)(a) of Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, such statements being likely strongly to dissuade certain candidates from submitting their candidature and, accordingly, to hinder their access to the labour market."

ruled that the statement delivered by a shareholder in a football club indicating that he would not hire openly homosexual footballers could amount to discrimination as a discriminatory hiring policy could be inferred from said statement.[688]

This is particularly relevant towards the issue at hand given that many algorithmic discrimination cases do not directly affect individuals. In this sense, it is possible to conclude that cases of targeted job advertising that explicitly exclude members of a protected group from seeing said advertisements could be considered to be covered by the Employment Directives even when no identifiable victims existed because the exclusion took place even before the recruitment process started.[689]

2.2. WHAT CONSTITUTES DIRECT ALGORITHMIC DISCRIMINATION?

For a case of algorithmic discrimination to fall under the scope of direct discrimination it would generally be necessary for the system to process a protected ground and consider it a relevant input factor in the decision at hand.[690] In the following paragraphs we establish several ways in which this process may take place.

The first and most obvious way in which this can happen is through the explicit introduction of protected grounds by algorithm designers. Firstly, programmers can specifically include protected attributes into the system alongside commands that specify the negative value of said grounds. However, it is not likely for this form of discrimination to be very common since explicitly introducing protected group membership data into the algorithm would automatically make it suspicious. Moreover, in some cases, regulatory instruments, such as the GDPR, forbid introducing protected grounds into automated decision-making systems. Hence, discriminatory outcomes will generally result from proxy variables for said protected grounds but will not exclusively affect members of the protected group and will have to be proven through the general impact that the decision has on the protected group, therefore constituting cases of indirect discrimination, which will be reviewed later on.[691]

[688] CJEU Judgment 25th April 2013, C-81/12, Asociaţia Accept v. Consiliul Naţional pentru Combaterea Discriminării.
[689] XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *cit.*, 2020, pp. 166-167.
[690] *Idem*, p. 169.
[691] *Idem*, p. 171.

In addition, considering the many possibilities for hiding discrimination that automated systems offer, it is not likely that tech experts who want to intentionally discriminate will decide to include protected grounds in the system.[692] Nonetheless, this form of explicit algorithmic discrimination could take place, for example, if programmers are confident that the system will remain opaque.

Secondly, it is also possible that protected categories are introduced in the system without initially providing them with a negative value and that the algorithm eventually learns to produce worse results for individuals who belong to specially protected groups.

Another case in which direct algorithmic discrimination will take place is when the system infers an individual's membership to an especially disadvantaged group from data to which it attributes a negative value and makes the decision either wholly or partly based on the (inferred) protected ground.[693] The algorithm's designers can articulate said inferences in a conscious or unconscious manner and it is also possible that the algorithm develops them once it is deployed. For these cases to fall under the scope of direct discrimination, the data from which protected group membership is inferred cannot be apparently neutral criteria, that is, they must not be significant in predicting the target variable.[694]

Amazon's sexist recruitment algorithm[695] constitutes a clear example of this type of direct algorithmic discrimination. The system was trained with the data from hiring decisions and CVs sent to the company over the previous ten years in order to determine the relevant characteristics that should be searched for in future applicants' résumés. Since the technology industry's workforce is mainly made up of men, the algorithm learned that sex was an important characteristic to search for in resumes. Although the algorithm did not have direct access to data related to job applicants' sex, it was able to infer said information from other elements. Thus, if a resume included a piece of information stating that the candidate had been "captain of women's chess club", the algorithm automatically lowered the applicant's score. In this case, the algorithm did not discriminate against women because it was instructed to do so, but because the algorithm had detected that "being a woman" was a piece

[692] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, pp. 712-714.
[693] DRECHSLER, L. & BENITO SÁNCHEZ, J. C., "The price is (not) right: data protection and discrimination in the age of pricing algorithms", *European Journal of Law and Technology*, vol. 9, No. 3, 2018, p. 13.
[694] GRIMMELMANN, J. y WESTREICH, D., "Incomprehensible discrimination", *California Law Review Online*, vol. 7, 2017, p. 176.
[695] See Chapter III. The target variable is the concept that the algorithm aims to measure, for example, "job performance" or "creditworthiness":

of information that it should look for since not being a woman was a typical characteristic of the people hired by the company during the previous ten years.[696]

It is therefore possible to establish two types of direct algorithmic discrimination: assigning a negative value to the specially protected category (explicit direct discrimination) or inference of the specially protected category through other pieces of data which are assigned negative values and which have no predictive value in the decision-making process (direct discrimination by inference). Within each of these two types of direct algorithmic discrimination it is also possible to establish two subtypes, depending on the way in which the protected characteristic is weighed in the system and the results generated by the algorithm and which can appear both in cases of explicit algorithmic direct discrimination and in cases of direct discrimination by inference.

The first subtype comprises those cases in which an individual's membership to a traditionally disadvantaged group automatically leads to the automated decision having negative effects for the individual. For example, if an algorithm detects that an individual belongs to an ethnic minority group, whether it is because it has direct access to that piece of information or because it infers it from other data, and it automatically denies her a loan as a result of her vulnerable group membership.

The second subtype includes instances of discrimination in which membership to the specially protected group is assigned a negative value that is them combined with the value that the individual obtains for other categories of data. For instance, in the automated system used by the Austrian public employment service, the general score that determines each individual's classification is established by combining the score obtained in each of the categories of data considered by the algorithm. The category "female" subtracts points but being an EU national increases an individuals' score.[697] Hence, it is possible that even when a discriminatory treatment towards members of specially protected groups exists in the decision, the final result does not discriminate against all members of disadvantaged groups if their other data points do not lower their score.

To sum up, direct algorithmic discrimination can take place in the following cases:

[696] BORNSTEIN, S., "Antidiscriminatory algorithms", *cit.*, 2019, p. 521; DASTIN, J., "Amazon scraps secret AI recruiting tool that showed bias against women", *cit.,* 10 de octubre de 2018.
[697] ALLHUTTER, D. *et al.*, "Algorithmic profiling of job seekers in Austria: how austerity politics are made effective", *Frontiers in Big Data*, vol. 3, 2020, pp. 1-17.

- Explicit direct discrimination in which the protected category determines the system's result: automatically harms all members of the protected group.
- Explicit direct discrimination in which the protected category is weighed with other variables: lowers score but does not fully determine final result.
- Inferred direct discrimination in which the protected category determines the system's result: automatically harms all members of the protected group.
- Inferred direct discrimination in which the protected category is weighed with other variables: lowers score but does not fully determine final result.

Finally, while intentionality does not play a significant role in the European anti-discrimination framework, it is also important to point out that all of the above stated types of direct algorithmic discrimination can be intentional or unintentional.

2.3. DETECTING AND PROVING DIRECT ALGORITHMIC DISCRIMINATION

Proving the existence of direct algorithmic discrimination will depend on a series of factors, which include, whether the system explicitly uses the protected variable, whether protected group membership determines the outcome or is weighed with other variables, whether the system is transparent and whether individuals have access to the information and knowledge of the way in which the decision has affected other members and non-members of the disadvantaged group. In this section, we establish and explain the factors at play and the different possible scenarios for detecting and proving direct algorithmic discrimination. However, before delving into said elements, the following paragraphs indicate and examine how direct discrimination can be measured as a function of individual fairness.

2.3.1. Direct algorithmic discrimination as a function of individual fairness

Individual fairness mechanisms define fairness from the perspective of the equal treatment of similarly situated individuals and are thus built on the idea "that any two individuals who are similar *with respect to a particular task* should be classified similarly."[698] In other words, the similarity of two individuals that is used as a measure of fairness must be considered with regard to the features that are relevant towards measuring the target variable. If two

[698] DWORK, C. *et al.*, "Fairness through awareness", *ITCS '12 Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012, p. 214, (emphasis in original).

individuals present similarities between said relevant features, they should receive a similar treatment.

This vision generally assumes that the training and test data are not skewed and that the way in which labels and features are measured and defined in the algorithm, that is, the way in which the algorithm measures reality, is mostly objective. Hence, only if there is a difference in outcome between similarly situated individuals who only differ in their protected characteristics, will the algorithm be considered unfair and said discrimination will be redressed.[699]

This version of fairness generally considers that suspect categories are not relevant towards predicting outcomes in most areas and, consequently, that algorithms that consider protected characteristics and offer different predictions depending on whether individuals are male or female, belong to a racial minority or not, etc., incur in instances of direct discrimination, the justification of which is only accepted by both the CJEU and ECHR in very specific cases and is applied in a very restrictive manner, as will be discussed later on.[700] Only in very few cases will it be possible to justify the use of a protected characteristic as a feature towards predicting the target variable.

2.3.2. When all members of the group are negatively affected by the decision

Cases in which the protected variable automatically determines a negative outcome for members of the group will be fairly easy to prove without having to access the source code. However, in this vein, only in cases where it is clear that all members of a specially protected group have been treated less favourably will it be easier for the courts to establish the existence of a case of direct discrimination without providing algorithmic transparency as it will be possible to apply, by analogy, the CJEU's ruling in the *Tadao Maruko* case,[701] which establishes that if the formal criterion by virtue of which the decision is made is inseparable from disadvantaged group membership, the action will constitute an instance of direct discrimination.

By analogy, an algorithm whose source code is unknown, but which always treats members of a certain specially protected group in a less favourable manner than non-members who are

[699] *Ibidem*.
[700] See section 2.4 in this chapter.
[701] CJEU Judgment 1st April 2008, C-267/06, Tadao Maruko v. Versorgungsanstalt der deutschen Bühnen.

in a substantially similar situation would be ruled as direct discrimination. This would be the case if, for example, all women were rejected in a recruitment process and only men were hired despite the fact that all candidates met the minimum requirements for the job. Thus, despite not knowing the element or elements that determine the decision, it is relatively easy to conclude that, if the decision results in less favourable treatment for all members of the group, it will have been made on the basis of the specially protected category or on elements of judgment that are inseparable from it. However, a certain level of transparency will also be necessary even in this case due to the fact that a plaintiff that suspects the decision to be discriminatory, will need access to information on the system's outcomes for a significant amount of people in order to prove that members of the disadvantaged group receive a less favourable treatment.

2.3.3. When not all members of the group are negatively affected by the decision

When the protected category is introduced into the algorithm and provided a negative value, the possibility of proving the existence of direct discrimination will be significantly hampered if not all members of the group suffer a negative consequence or if an average cannot be drawn showing that members of the disadvantaged group are treated less favourably than non-members. Thus, when the algorithm attributes a negative value to the membership of a disadvantaged group but the final result does not depend exclusively on this characteristic but on a combination of variables, proving direct discrimination will not be easy without transparency. Hence, in principle, the applicability of prohibitions of direct discrimination is considerably reduced in these cases.

Moreover, whenever direct discrimination results from an algorithm inferring protected group membership and then said data is weighed with other variables, lowering individuals' scores but not determining the final result, it will be necessary to access the system to examine whether the variables that are correlated with the protected category are or not predictive of the target variable, say for instance, "job performance", or are simply serving a self-fulfilling prophecy in which the algorithm has learned that "poor job performance" actually means not being white.

Therefore, in principle, in order to prove a case of algorithmic discrimination, transparency will be needed. Some cases will only require sufficient transparency to access the results for other individuals subjected to the same decision-making process with the objective of

providing a comparator. However, others may require full algorithmic transparency. While there are no specific cases regarding the need to disclose algorithms in the context of discriminatory decisions, the *Danfoss*[702] and *Galina Meister*[703] rulings offer useful insights into the ways in which the CJEU could rule with regard to algorithmic transparency in the context of alleged discrimination cases.

In *Danfoss*, the CJEU established that:

> "Where an undertaking applies a system of pay which is totally lacking in transparency, it is for the employer to prove that his practice in the matter of wages is not discriminatory, if a female worker establishes, in relation to a relatively large number of employees, that the average pay for women is less than that for men…"

Conversely, the *Galina Meister* ruling establishes that the provisions contained in the EU Equality Directives setting the rules for the burden of proof in discrimination cases "must be interpreted as not entitling a worker who claims plausibly that he meets the requirements listed in a job advertisement and whose application was rejected to have access to information indicating whether the employer engaged another applicant at the end of the recruitment process."

While the Court does burden the respondent with proving that an opaque pay system is not discriminatory, it only does so once it has been proven that said system yields worse results on average for women than it does for men. However, if the plaintiff only suspects that the decision was discriminatory but cannot provide an element of comparison or proof of discriminatory treatment, the respondent has no obligation to disclose the underlying logic to the decision. Following from this reasoning, if an algorithm yields discriminatory results but the individuals affected are unable to prove discriminatory treatment because they do not have access to the results produced for other individuals, the respondent will not have to provide an explanation or make the system transparent.

While it is also important to consider that, as it will be addressed to a further extent in part II, several courts and administrative bodies have established the need to set full system

[702] CJEU Judgment, 17th October 1989, C-109/88, Union of Commercial and Clerical Employees, Denmark v. Danfoss A/S.
[703] CJEU April 19th April 2012, C-415/10, Galina Meister v. Speech Design Carrier Systems GmbH.

transparency when it comes to public sector automated decision-making,[704] even if this trend is adopted when it comes to the evaluation of discrimination cases, this does not mean that transparency requirements will increase. Firstly, the question of whether said transparency requirements will also apply to private parties still remains. Secondly, if Courts keep requiring individuals to provide a comparator in most cases, and therefore prove that there has been an instance of differential treatment, in order for the algorithm to be disclosed, the chances that plaintiffs have in presenting successful requests for transparency may be hindered if the data of other individuals subjected to the same decision-making process is considered to fall under the scope of personal data or intellectual property rights. Nevertheless, this problem may be solved by requesting aggregate anonymised data that only identifies and differentiates between the effects of the algorithm on members and non-members of the protected group.[705]

Additionally, even when individuals have access to the results of other individuals subjected to the same algorithmic decision, depending on how automated decision-making is interpreted by Courts, it may lead either to enhancing or hampering the chances of providing a valid comparator. On the one hand, automated decision-making systems process the data of many individuals and the decisions they yield can affect very large groups of people. Thus, in theory, it is easier to provide a comparator. However, since machine learning systems are continuously developing and are sometimes able to provide specific and quasi-personalised decisions for each individual, it is also possible for Courts to interpret that there are no available comparators for the decision process an individual has been subjected to. Another related problem that appears with regard to the ever-evolving nature of automated systems is that Courts must judge cases based on the software's version that was used in the decision-making process, which might not longer be available at the time in which the case is judged. If that software's version has not been saved, Courts will not even have access to the discriminatory measure.

For cases in which transparency is not granted or is not possible due to the system's complex nature, several scholars have put forward proposals for algorithmic accountability without

[704] Administrative Regional Court of Lazio-Roma, Section III bis, Judgments No. 3769, 22nd March 2017 and No. 10964, 13th September 2019; Catalan Commission for the Guarantee of the right of access to public information, Joined decisions 123/2016 and 124/2016; and, French Commission on access to administrative documents, decisions No. 20144578 of 8th January 2015 and No. 20180276 of 19th April 2018

[705] HACKER, P., "Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law", *Common Market Law Review*, vol. 55, No. 4, 2018, pp. 1179, 1182.

opening the "black box"[706] that could also be incorporated to anti-discrimination judicial procedures in order to prove the existence of discriminatory treatment. Said proposals generally include forcing controllers to provide some information regarding the way in which the algorithm is designed, including the training data, test data and how the different labels and target variable are defined.[707] By providing this information and the actual programme, plaintiffs could tinker with the system in order to determine the effects that protected group membership has on outcomes,[708] hence establishing hypothetical comparators through experimentation with the system. In order for this system to work in algorithms that keep evolving after they have been deployed, it will be necessary for each version of the software used in decision-making processes to be saved.

This system will probably work quite well for algorithms that are not blind to protected group membership, since, for example, in scoring algorithms, once the protected characteristic is introduced (keeping all other variables constant), it will be possible to determine whether the total score is reduced. KLEINBERG *et al.* also argue that this system will be effective regardless of whether protected group membership is included or whether the algorithm is blinded to said pieces of data.[709] However, it will be much harder to prove cases of direct discrimination in which the algorithm does not automatically determine a negative consequence for all members of the group but which, for example, provides a lower score to certain characteristics from which protected group membership is inferred but that are not predictive of the target variable. In these cases, it is unlikely that the system based on counterfactuals is be able to prove the lack of predictive value of said characteristics for they will not be known unless the black box is opened. These cases will probably have to be presented and defended as instances of indirect discrimination. Considering the wider scope for justifying cases of indirect discrimination, that will be reviewed in the following section, this necessary shift in strategy can cause significant harms to the rights to equality and non-discrimination. Nonetheless, considering the degree at which machine learning technologies are being developed, we do not rule out that, systems that, when being operated by the

[706] PASQUALE, F., *The Black Box Society*…, *cit.*, 2015.
[707] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, pp. 1085-1139; KLEINBERG, J. *et al.*, "Discrimination in the age of algorithms", *Journal of Legal Analysis*, vol. 10, 2018, pp. 113-174; WACHTER, S., MITTELSTADT, B. D. & RUSSELL, C., "Counterfactual explanations without opening the black box: automated decisions and the GDPR", *Harvard Journal of Law & Technology*, vol. 31, No. 2, 2018, pp. 841-887.
[708] KLEINBERG, J. *et al.*, "Discrimination in the age of algorithms", *cit.*, 2018, p. 150.
[709] *Ibidem.*

relevant specialists, can effectively detect most forms of algorithmic direct discrimination through counterfactuals are soon set up.

This section has conveyed the possibilities of proving the different instances of algorithmic discrimination that fall within the scope of direct discrimination or disparate treatment. While it is possible to prove direct algorithmic discrimination, in many cases it will require resorting to the application of other technological elements and experts. Thus, unless specific protocols are established for judging cases of algorithmic discrimination, it is highly likely that most of them will be judged as instances of indirect discrimination, thus focusing on the outcome rather on the actual treatment.

2.4. JUSTIFICATIONS TO DIRECT ALGORITHMIC DISCRIMINATION

Within the EU framework, the justifications admitted by the CJEU and the ECHR in cases of direct discrimination vary slightly. The ECHR refers to a set of requirements that must be met in order for a certain action or lack thereof to be considered discriminatory. The elements that must be evaluated are "whether (1) there is differential treatment of (2) equal cases; whether there is (3) an objective and reasonable justification; and, whether there is (4) proportionality between aim and means".[710]

Since the aforementioned elements apply to both direct and indirect discrimination, the ECHR therefore admits the employment of an objective and reasonable justification for both types of discrimination. In general, for the Court to admit the validity of the justification provided by defendants, they have to prove that there is no other, less aggressive measure, in order to achieve the objective and that the objective is sufficiently important to justify the level of difference in treatment.[711] However, the Court requires "particular weighty reasons to justify discrimination on grounds of sex, sexual orientation, race, colour, nationality (except in regard to immigration), illegitimacy and religion".[712] In fact, when analysing cases of differential treatment on the basis of gender, the court has generally only ruled in favour of

[710] VANDENHOLE, W., *Non-discrimination and Equality in the View of the UN Human Rights Treaty Bodies*, Antwerpen – Oxford, Intersentia, 2005, p. 34.
[711] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", *cit.*, 2018, p. 93.
[712] COUNCIL OF EUROPE, "Prohibition of discrimination", 2017. Available on 5th June 2019 at: https://www.coe.int/

said treatment when it benefitted women over men in order to correct for historical and persisting situations of discrimination.[713]

Conversely, the CJEU only admits justifying direct discrimination in cases that fall under the scope of the exceptions set by the Equality Directives. The CJEU has generally interpreted said exceptions very strictly.

Firstly, the Directives on equality in employment establish that applying differential treatment on the basis of one of the protected categories "shall not constitute discrimination where, by reason of the nature of the particular occupational activities concerned or of the context in which they are carried out, such a characteristic constitutes a genuine and determining occupational requirement, provided that the objective is legitimate and the requirement is proportionate."[714] Following this exception it is possible to set maximum age limitations for hiring employees for specific jobs that are especially physically demanding.[715] Additionally, the Court has admitted that there are certain professions, such as those developed in artistic sectors, for which the selection of an individual based on protected grounds might be required.[716]

However, since the justifications for differential treatment have been must in all cases be interpreted restrictively,[717] the Court has for example considered that customers' preference to not be served by women that wear Islamic headscarves could not be deemed "a genuine and determining occupational requirement" that could justify the termination of a contract as it referred to the subjective considerations of the employer, who decided to take customer's wishes into account.[718]

---

[713] See, for example, ECHR Judgments 17th February 2011, 6268/08, Andrle v. Czech Republic; 10th May 2007, 42949/98 and 53134/99, Runkee and White v. United Kingdom; 24th January 2017, 60367/08, Khamtokhu and Aksenchik v. Russia.

[714] Gender Employment Equality Directive, Art. 14 (2); Racial Equality Directive, Art. 4; Employment Equality Directive, Art. 4 (1).

[715] CJEU Judgment 12th January 2010, C-229/08, Colin Wolf v. Stadt Frankfurt am Main, paragraph 40.

[716] CJEU Judgment 21st May 1985, C-248/83, Commission of the European Communities v. Federal Republic of Germany.

[717] CJEU Judgment 15th May 1986, C-222/84, Johnston v. Chief Constable of the Royal Ulster Constabulary, paragraph 36.

[718] CJEU Judgment 14th March 2017, C-188/15, Asma Bougnaoui and Association de défense des droits de l'homme (ADDH) v. Micropole SA, paragraph 40: "…the concept of a 'genuine and determining occupational requirement', within the meaning of that provision, refers to a requirement that is objectively dictated by the nature of the occupational activities concerned or of the context in which they are carried out. It cannot, however, cover subjective considerations, such as the willingness of the employer to take account of the particular wishes of the customer."

This particular ruling could be applied by analogy to the massive power that clients hold in sharing economy platforms. The existence of a gender pay gap in sharing economy platforms partly results from the stereotypes that customers have which then lead to assigning lower ratings to female platform workers.[719] Moreover, low ratings do not just lead to lower wages but can even mean that drivers are expelled from sharing economy platforms.[720] Hence, client preferences reflected in the ratings they provide should not affect working conditions and pay.[721] However, as it will be discussed later on, in order to apply this particular case law to platform workers, they should be considered employees and not self-employed under EU law.[722]

Secondly, the Employment Equality Directive recognises exceptions to the principle of equality within churches and other organisations as well as with respect to setting age limits for organisational purposes. With regard to the former, the Directive enables said organisations to require individuals working for them to act in good faith and with loyalty to the organisation's ethos, as long as said policy falls within the boundaries of national constitutions and laws. The CJEU has limited the scope of justification of this particular type of religious discrimination and considered that, only if the religious requirement constitutes an objective requirement necessary for the development of the occupational activity, will discrimination be allowed. While the ethos and right of autonomy of the religious organisation in question must be considered in such cases, it is possible for courts to interpret and determine whether an occupational requirement that discriminates on the basis of religion is objective and legitimate, thus limiting the possibility that religious organisations establish said requirements with absolute freedom.[723]

As for the age exception, the Directive states that differences in treatment may include the fixing of a maximum age for recruitment which is based on the training requirements of the

[719] RENAN BARZILAY, A. & BEN-DAVID, A., "Platform inequality: gender in the gig economy", *Seton Hall Law Review,* vol. 47, No. 393, 2017, p. 404; XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *cit*., 2020, pp. 160-161. Other studies have shown the gender pay gap in transport platforms, such as Uber, to mainly be the result of driver "preferences". See COOK, C. *et al.*, "The gender earnings gap in the gig economy: evidence from over a million rideshare drivers", *NBER* Working Paper No. w24732, 2019, pp. 1-62. Available on 2nd March 2020 at: https://www.nber.org/

[720] KÜLLMANN, M., "Platform work, algorithmic decision-making, and eu gender equality law", *cit.*, 2018, p. 8.

[721] XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *cit*., 2020, p. 161.

[722] KÜLLMANN, M., "Platform work, algorithmic decision-making, and eu gender equality law", *cit.*, 2018, p. 12.

[723] Art. 4.1 Employment Equality Directive. See CJEU Judgment 17th April 2018, C-414/16, Vera Egenberger v. Evangelisches Werk für Diakonie und Entwicklung e.V for a recen interpretation of the way in which member state Acts implementing the Employment Equality Directive should include this justification to direct discrimination.

post in question or the need for a reasonable period of employment before retirement. In a similar vein, the CJEU has stated the possibility of setting a compulsory age of retirement with "the aim of putting in place a balanced age structure in order to facilitate planning of staff departures, ensure the promotion of civil servants, particularly the younger ones among them, and prevent disputes that might arise on retirement".[724]

Moreover, article 2.5 of the Employment Equality Directive allows member states to establish exceptions on the basis of "public security, for the maintenance of public order and the prevention of criminal offences, for the protection of health and for the protection of the rights and freedoms of others."

While justifications to cases of direct discrimination have only been admitted in very exceptional cases both by the ECHR and CJEU,[725] the apparent, and sometimes real, accuracy of automated systems could lead Courts to become much more lenient when accepting the justifications provided by respondents as being valid.

2.5. PROBLEMATIC CASES

2.5.1. Accurate direct algorithmic discrimination

From an economic perspective, algorithmic discrimination can generally be described as statistical discrimination. In the case of direct algorithmic discrimination, the automated system uses the protected characteristic as an element that is indicative of other quality. For example, race may be used as a proxy for past criminal history.[726]

Associating a certain negative characteristic with a protected group and assuming that said feature can be attributed to all members of the group, is not just wrong because it will perpetuate structural discrimination and disadvantage but also because it is unfair at an individual level as it can lead to erroneously profiling individual members of said group who

[724] CJEU Judgment 21st July 2011, C-159/10 and C-160/10, Gerhard Fuchs and Peter Köhler v. Land Hessen, paragraph 60.
[725] CJEU Judgment 15th May 1986, C-222/84, Johnston v. Chief Constable of the Royal Ulster Constabulary, paragraph 36; CJEU Judgment 3rd February 2000, C-207/98, Mahlburg v. Land Mecklenburg-Vorpommern; CJEU Judgment 12th January 2010, C-341/08, Domnica Petersen v. Berufungsausschuss für Zahnärzte für den Bezirk Westfalen-Lippe, paragraphs 60-64; CJEU 13th September 2011, C-447/09, Reinhard Prigge and Others v. Deutsche Lufthansa AG.
[726] DOLEAC, J. L. & HANSEN, B., "The unintended consequences of 'ban the box': statistical discrimination and employment outcomes when criminal histories are hidden", *Journal of Labour Economics*, vol. 38, No. 2, 2020, pp. 321-374; STRAHILEVITZ, L., "Privacy versus antidiscrimination", *cit.*, 2008, pp. 363-381.

do not fit the stereotype.[727] Unfortunately, false positives and negatives almost always come up whenever proxies are used in order to identify and measure the features that are actually relevant for decision-makers. It is therefore necessary to measure and determine the extent to which we can accept the trade-off between ensuring that all individual decisions are fair (based on real data) and the efficiency produced by using certain shortcuts, such as proxies or correlations that do not entail causality and may therefore wrongfully attribute certain characteristics to individuals.

In this sense, the *Test-Achats*[728] CJEU judgment declared that altering insurance premiums on the basis of gender, as initially allowed by article 5.2 of the Gender Goods and Services Directive, was contrary to articles 21 and 23 of the Charter of Fundamental Rights of the EU and articles 2, 3 and 6 of the Treaty of the European Union. This ruling mainly benefits a group that is not structurally disadvantaged (but that is disadvantaged in this specific area), seeing as it is men who are generally assigned higher insurance premiums due to their shorter life expectancy and more aggressive driving attitudes, amongst other elements. The judgment is nonetheless relevant as it set an important precedent regarding prohibitions to base discriminatory treatment on statistical data that is accurate for the group in general but might not be so for specific individuals.[729]

2.5.2. Direct discrimination by association

According to EU case law, direct discrimination also takes place when a protected ground does not concur in the individual who is directly harmed by a provision, criterion or practice but suffers negative consequences due to her relationship with members of a protected group. In the *Coleman*[730] case, the Court determined that the carer (mother) of a person with disability had been harassed and suffered other forms of discrimination by her employer as a consequence of the special needs of her child and concluded that "the prohibition of direct

[727] MITTELSTADT, B. D., "The ethics of algorithms…", *cit.*, 2016, p. 5.
XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *cit*., 2020, p. 157.
[728] CJEU Judgment 1st March 2011, C-236/09, Association belge des Consommateurs Test-Achats ASBL, Yann van Vugt, Charles Basselier v. Conseil des ministres.
[729] Opinion of Advocate General Kokott delivered on 30th September on Case C-236/09, paragraph 62: "…many other factors play an important role in the evaluation of the abovementioned insurance risks. Thus, for instance, the life expectancy of insured persons, which is of particular interest in the present case, is strongly influenced by economic and social conditions as well as by the habits of each individual (for example, the kind and extent of the professional activity carried out, the family and social environment, eating habits, consumption of stimulants and/or drugs, leisure activities and sporting activities)."
[730] CJEU Judgment July 17th 2008, C-303/06, S. Coleman v. Attridge Law and Steve Law.

discrimination … is not limited only to people who are themselves disabled" and that the "less favourable treatment of that employee is based on the disability of his child, whose care is provided primarily by that employee … is contrary to the prohibition of direct discrimination laid down by Article 2(2)(a)"[731] of the Employment Equality Directive.

Hence, direct discrimination by association also falls within the scope of the right to equality and non-discrimination and the specific protections it offers to disadvantaged groups,[732] thereby also offering the possibility of redirecting discrimination by perception cases through the channels of special protection offered to members of disadvantaged groups.[733] Recognising such cases as deserving of the special protections offered to members of disadvantaged groups is highly relevant in the context of erroneous algorithmic profiling. An individual who is not a member of the discriminated group but who is erroneously identified as such and discriminated against due to said inaccurate classification through profiling techniques could therefore bring a direct discrimination claim against said action.[734]

In addition, offering non-members of protected groups the possibility of claiming discriminatory treatment on prohibited grounds allows for members of the protected group who might not want to publicly state that they are members of the disadvantaged group, to bring about the claim under the umbrella offered by affinity profiling, without having to identify as members of the group.[735]

In this sense, the use of profiling algorithms by Facebook in order to offer advertisers the possibility of excluding individuals associated with certain ethnicities could fall under this form of direct discrimination.[736]

2.6. HARASSMENT

Harassment is a form of direct discrimination that has been treated separately due to its specific characteristics. Harassment takes place "where unwanted conduct related to any of

[731] *Idem*, paragraph 56.
[732] WACHTER, S., "Affinity profiling and discrimination by association in online behavioural advertising", *Berkeley Technology Law Journal*, vol. 35, No. 2, 2020 (forthcoming), pp. 32-36. Available on 15th March 2020 at: https://papers.ssrn.com/
[733] *Idem*, p. 35.
[734] *Ibidem*.
[735] *Idem*, p. 36.
[736] WACHTER, S., "Affinity profiling and discrimination by association in online behavioural advertising", *cit.* 2020 (forthcoming), pp. 32-36; ANGWIN, J. & PARRIS JR., T., "Facebook lets advertisers exclude users by race", *cit.*, 2016.

the [protected grounds] takes place with the purpose or effect of violating the dignity of a person and of creating an intimidating, hostile, degrading, humiliating or offensive environment".[737] Although it may seem difficult to bring about algorithmic discrimination claims on the basis of harassment, the EU framework takes into consideration the victim's own perception of said treatment. It is therefore possible that certain forms of algorithmic discrimination that take place through the perpetuation of stereotypes that negatively affect protected groups are brought before the courts as cases of harassment.[738]

For national and supranational courts to consider these cases as instances of harassment, they will have to accept the possibility that there is no one single individual who these systems have harmed and, consequently, no specific victim status, but the existence of a more general and indirect harm caused on the disadvantaged group as a whole. In these cases, the absence of victim status goes beyond what is currently accepted by CJEU case law seeing as there is no announcement of actions to discriminate as there was in the *Feryn* and *Asociaţia Accept*[739] cases, but actions or expressions that help to perpetuate stereotypes and, consequently, structures of disadvantage. Hence, unless a system of collective action streamed through associations that are representative of the interests of disadvantaged groups is not implemented, it seems highly unlikely that the CJEU and, specially, the ECHR will consider harassment claims stemming from uses of algorithms in search engines or personal assistants.

3. INDIRECT ALGORITHMIC DISCRIMINATION

Indirect discrimination takes place when an apparently objective rule results in discriminatory outcomes for one particular sex, race, sexual orientation, religion, etc.[740] Indirect discrimination therefore does not focus on the treatment provided by these decisions, which is apparently neutral, but on the actual impact they have on the group as a whole.

---

[737] Racial Equality Directive, Art. 2.3; Employment Equality Directive, Art. 2.3; Gender Goods and Services Directive, Art. 2.c; Gender Employment Equality Directive, Art. 2.1.c.

[738] ALLEN, R. & MASTERS, D., "Artificial Intelligence…", *cit.*, 2020, p. 593: "There is plenty of scope for technology to lead to harassment. In August 2016, Snapchat introduced a face-morphing filter which was "inspired by anime". In fact, the filter turned its users' faces into offensive caricatures of Asian stereotypes. Smart phone assistants nearly all have default female voices e.g. Apple's Siri, Google Now and Microsoft's Cortana. This echoes the dangerous gender stereotype that women, rather than men, are expected to be helpful and subservient".

[739] CJEU Judgments 10th July 2008, C-54/07, Firma Feryn NV v. Centrum voor gelijkheid van kansen en voor racismebestrijding and 25th April 2013, C-81/12, Asociaţia Accept v. Consiliul Naţional pentru Combaterea Discriminării

[740] XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *cit.*, 2020, p. 170.

This notion of discrimination was conceived in order to fill the gaps left in formal approaches to equality and discrimination.[741] Unlike direct discrimination, the construction of indirect discrimination brings about material considerations and does not exclusively rely on formal elements. The recognition of indirect discrimination entails acknowledging the existence of unequal points of departure between members of dominant groups and members of disadvantaged groups which limits the actual freedom of the latter. This concept does, however, still take up the basic elements offered by notions of formal equality and addresses instances of discrimination in a static manner and not necessarily as the result of a history of subordination. Indirect discrimination takes place when different criteria are applied to different groups that, although not formally segregated by a protected characteristic, are factually differentiated by a protected characteristic. Hence, while assessing the case of discrimination through the apparently neutral criterion's impact, indirect discrimination also covers, and aims to deal with, unequal treatment.[742]

This form of discrimination does provide a vital element in order to properly address discrimination cases as it recognises the group or collective dimension of discrimination. The US Supreme Court's landmark 1971 *Griggs v. Duke Power Co.* ruling[743] was fundamental for the development of a theory of discrimination applicable to apparently neutral situations which hid instances of discrimination, which in the US was labelled "disparate impact" and is generally equivalent to indirect discrimination. In said case, the US Supreme Court ruled that requiring job or promotion candidates to be in possession of a high school diploma and to be subjected to intelligence testing was discriminatory against African Americans given the fact that they generally had more difficulties in fulfilling said requirements. The Court ruled that only if the educational degree and test were necessary in order to develop the job would they be deemed justified.

This ruling constitutes a milestone in the development of anti-discrimination law enforcement given the fact that it considered the possibility of discriminatory situations occurring outside

[741] MORRIS, A. J., "On the normative foundations of indirect discrimination law: understanding the competing models of discrimination law as Aristotelian forms of justice", *Oxford Journal of Legal Studies*, vol. 15, No. 2, 1995, p. 202: "The individual justice model … was the sole or dominant theory of anti-discrimination law until the mid-1970s, when dissatisfaction with its limitations began to develop. Writers argued, for instance, that it fails to account for the effects of history and for the deep institutional nature of discrimination and to 'take[] full[] account of social reality'. From these criticisms emerged a generally opposing set of ideas, called by Fallon and Weiler the model of group justice. Its animating notion is to 'protect and advance the interests of historically disadvantaged groups, especially blacks'."
[742] AGUILERA RULL, A., *Contratación y Diferencia*…, *cit.*, 2013, pp. 177-178.
[743] US Supreme Court, *Griggs v. Duke Power Co.*, 401 U.S. 424, 1971.

of the framework of cases in which the protected ground was directly linked to the discriminatory action, therefore considering the group-dimension and structural origin of instances of discrimination.[744] The introduction of the concept of indirect discrimination in the European regulatory framework was carried out by the CJEU's Judgment in the *Defrenne II*[745] case but it was done so in a fragmented and confusing manner[746] as it referred to indirect discrimination in the following terms:

> "Direct and overt discrimination which may be identified solely with the aid of the criteria based on equal work and equal pay referred to by the article in question and, secondly, indirect and disguised discrimination which can only be identified by reference to more explicit implementing provisions of a community or national character."[747]

Thus the court, did not, in this case, establish the definition of indirect discrimination as it is currently widely accepted but referred to this type of this discrimination as disguised discrimination. However, the importance of this judgment must not be downplayed as it introduced the concept of indirect discrimination to the European legal system.

3.1. ESTABLISHING A 'PRIMA FACIE' CASE OF INDIRECT DISCRIMINATION

Following the definitions provided by the EU Directives on equality and non-discrimination, most EU member states have included prohibitions regarding indirect discrimination. It is important, to note that, while non-member states do generally include general prohibitions against discrimination, it is not as common to find specific references to indirect discrimination as it is within EU member state legislation.[748]

The first step in proving indirect discrimination is establishing a *prima facie* case of indirect discrimination. Once this is accomplished, the respondent will have to prove that the

[744] BARRÈRE UNZUETA, M. A., *Discriminación, Derecho Antidiscriminatorio y Acción Positiva a favor de las Mujeres*, Madrid, Civitas, 1997, pp. 43.
[745] CJEU Judgment 8th April 1976, C-43/75, Gabrielle Defrenne v. Société anonyme belge de navigation aérienne Sabena..
[746] BARRÈRE UNZUETA, M. A., "Problemas del derecho antidiscriminatorio…", *cit*., 2003a, p. 7.
[747] CJEU Judgment 8th April 1976, C-43/75, Gabrielle Defrenne v. Société anonyme belge de navigation aérienne Sabena, paragraph 18.
[748] CHOPIN, I. & GERMAIN, C., "A comparative analysis of non-discrimination law in Europe 2019", *cit.*, 2019, p. 47.

principle of equal treatment has not been breached.[749] For a *prima facie* case of indirect discrimination to be established plaintiffs must provide evidence of the existence of (1) an apparently neutral criterion (2) that results in more negative effects for members of the protected group (3) when compared to non-members.

3.1.1. An apparently neutral criterion that results in more negative effects for members of the protected group

The first element is the existence of a neutral rule that is applied to everyone irrespective of each individual's oppressed group membership.[750] An example of a neutral criterion would be a minimum general height requirement of 1,70m to access a certain occupation. The second element that is required is that the apparently neutral criterion results in more negative effects for members of the protected group.[751] Following from the previous example, given the fact that men are on average taller, would result in women being negatively affected by this measure.

With regard to these two first criteria, it is important to highlight that indirect discrimination in many instances results from the lack of consideration of the specific experiences and attributes of disadvantaged groups. In this sense, an especially interesting type of indirect discrimination is the one that results from the hierarchisation of values and characteristics that are assigned to dominant and groups that have historically suffered oppression, for example, leading to differences in pay between typically male and typically female occupations.

While the main focus of this chapter is the European anti-discrimination framework, it is relevant to cite the Spanish Constitutional Court's landmark Judgment No. 145/1991 which clearly illustrates the way in which considering the superiority of typically male characteristics leads to the undervaluation of occupations that are mostly developed by women. In said judgment, the Court considered a case in which the cleaners (a professional group mostly occupied by women) earned less than the labourers (a professional group mostly occupied by men) in a Spanish hospital even though they carried out identical tasks. The relevance of this decision however mostly lies on the ruling concerning not the particular

[749] See art. 8 of the Racial Equality Directive, art. 10 of the Employment Equality Directive, art. 19 of the Gender Employment Equality Directive and art. 9 of the Gender Goods and Services Directive.
[750] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", *cit.*, 2018, p. 54.
[751] *Idem*, p. 55.

case of workers that developed identical tasks, which was deemed a case of direct discrimination,[752] but of the collective employment agreement that applied to all workers of the Region of Madrid's public health sector.

In the referred collective employment agreement, labourers were generally assigned higher retributions due to the fact that the work they carried out, which according to the agreement was in fact different from the tasks developed by cleaners, required a greater physical effort than the work carried out by workers hired under the professional category "cleaners". The Spanish Constitutional Court established in said judgment the principle of "equal pay for work of equal value" by indicating that the constitutional principle of non-discrimination in relation to pay also covers all those cases in which there is an unequal attribution of value to jobs that are not strictly equal but equivalent or of equal value, when the determining factor for the unequal attribution of value is worker's sex or elements that are linked to it. The ruling also indicated that the principle of non-discrimination in relation to pay particularly excludes the consideration of female status when attributing value to work, for it reflects the social and economic undervaluation of women's work.[753]

Specifically, with regard to the use of "physical effort" as a parameter according to which the value of work was established, the Court indicated that:

> "…the exclusive and unreasonable use of this objective criterion has produced unequal and harmful effects for women. It takes as a point of departure an assumption that has not been proven, the greater unpleasantness and physical effort, attributing greater value to a predominantly male quality, ignoring other characteristics of work (attentiveness, care, assiduity, responsibility, etc.) that are more neutral with regard to their impact on each sex."[754]

The CJEU has also highlighted the need to attribute equal conditions to work of equal value in several judgments, indicating the importance of employing neutral parameters when

[752] Spanish Constitutional Court Judgment No. 145/1991, Section 3.
[753] *Idem*, Section 4.
[754] *Idem*, Section 5.

carrying out assessments regarding a particular job's value. For example in Case C-400/93 Royal Copenhagen[755] it ruled as follows:

> "32 Consideration of whether the principle of equal pay has been observed requires a comparison between the pay of workers of different sexes for the same work or for work to which equal value is attributed.
>
> 33 Where such a comparison involves the average pay of two groups of workers paid by the piece, it must in order to be relevant encompass groups each comprising all the workers who, taking account of a set of factors such as the nature of the work, the training requirements and the working conditions, can be considered to be in a comparable situation."[756]

The undervaluation of the specific characteristics of traditionally oppressed groups by developing assessments structured upon apparently neutral parameters which in fact prioritise and ascribe greater value to the attributes associated with groups that are generally in a position of social advantage, is common not only with regard to work but to many other aspects of society. However, illustrating this idea from the perspective of the pay gap, and more specifically given the cited cases, the gender pay gap, helps to convey the way in which said apparently neutral assessments lead to situations of tangible material inequality, thereby reinforcing historical situations of discrimination against traditionally disadvantaged groups.

As it was previously stated, since algorithms will not generally contain protected grounds as input variables that will influence the decision-making process, the element or elements that result in a discriminatory decision will not be directly visible and will therefore seem neutral while in fact correlating with a protected group and thus resulting in a negative impact for groups that are mostly or significantly made up of members of protected groups, such as the use of postal codes to determine loan eligibility.[757]

---

[755] CJEU Judgment 31st May 1995, C-400/93, Specialarbejderforbundet i Danmark v. Dansk Industri, formerly Industriens Arbejdsgivere, acting for Royal Copenhagen A/S.

[756] Similar rulings can be found in CJEU Judgments 28th February 2013, C-427/11, Margaret Kenny and others v. Minister for Justice, Equality and Law Reform, Minister for Finance, Commissioner of An Garda Síochána, paragraph 27; 26th June 2001, C-381-99, Susanna Brunnhofer v. Bank der österreichischen Postsparkasse AG, paragraph 43; 11th May 1999, C-309/97, Angestelltenbetriebsrat der Wiener Gebietskrankenkasse v. Wiener Gebietskrankenkasse, paragraph 17.

[757] ALLEN, R. & MASTERS, D., "Artificial Intelligence…", *cit.,* 2020, p. 592; HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, pp. 1143-1186.

As it was indicated in the previous chapter, the existence of biased variables that lead to algorithmic discrimination in many cases results from the way in which problems are specified and labels are structured. Since many of the phenomena measured by automated systems are not objective, the decisions made regarding the way in which the predictive goal should be shaped unavoidably contain certain degree of subjectivity.[758] If problems are specified and labels are defined in a way that disadvantages the values or elements associated with protected groups, they will lead to the discrimination of members of said groups.[759]

Other cases in which apparently neutral algorithmic decisions will result in discriminatory impact are those in which the negative impact on protected groups results from under or over-representation in the dataset or from the reproduction of structural discrimination.[760] Stereotyping is also a mechanism that can lead to cases of indirect discrimination and is closely related to value hierarchisation.[761] This pathway to discrimination was already discussed in the previous section, referred to direct discrimination, seeing as stereotyping can mean directly using a protected ground as a significant variable to predict individuals' behaviour. However, stereotypes can also be used in an unconscious or indirect manner when a characteristic is assigned to an individual because of their protected group membership but said relationship between assigning a feature and the protected ground is not explicitly stated.

3.1.2. A comparator

The third and final element that is required to prove the existence of indirect discrimination is a comparator. The disadvantaged group must be disadvantaged with respect to the impact that the same measure has on another (advantaged) group. It is important to highlight that the disadvantaged group will generally not be exclusively made up of members of a disadvantaged group. An example of this is the case of *Isabel Elbal Moreno v. Instituto Nacional de la Seguridad Social, Tesorería General de la Seguridad Social*.[762] In said case the CJEU concluded that the provisions set by the Spanish General Law on Social Security which required a proportionally greater contribution from part-time workers than from full time workers in order for the former to access retirement pensions was contrary to the

[758] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, pp. 678-680.
[759] *Idem*, p. 680.
[760] BORNSTEIN, S., "Antidiscriminatory algorithms", *cit.*, 2019, pp. 526.
[761] XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *cit*., 2020, pp. 156-157.
[762] CJEU Judgment 22nd November 2012, C-385/11, Isabel Elbal Moreno v. Instituto Nacional de la Seguridad Social, Tesorería General de la Seguridad Social.

prohibition of discrimination set by article 4 of the Directive on the progressive implementation of the principle of equal treatment for men and women in matters of social security given the fact that 80% of part-time workers in Spain were women. Consequently, even though the apparently neutral measure did not result in a disadvantage exclusively for women, since it mainly disadvantaged them it was ruled to be a form of indirect discrimination.

As with direct discrimination cases, providing a comparator in order to prove instances of indirect discrimination is also problematic not just in algorithmic decision-making but also with regard to traditional forms of discrimination, seeing as full awareness of the way in which a decision has impacted the different subjects in the decision-making process can sometimes prove highly difficult. In addition, once an individual has realised she has been subjected to a discriminatory algorithmic decision, it is also possible that the information necessary to provide a comparator might not be available if it falls under intellectual property rights or within the scope of personal data or that the comparator is not deemed acceptable by the Court as a result of the constant and self-evolving nature of algorithms.

3.2. DETERMINING AND PROVING *PRIMA FACIE* INDIRECT ALGORITHMIC DISCRIMINATION

3.2.1. General considerations

Unlike direct discrimination cases, indirect discrimination deals with the discriminatory impact of the decision. Hence, proving that the decision was based on the protected ground is irrelevant. It is therefore likely that many cases in which the algorithm is directly discriminating against members of disadvantaged groups by inferring said data from other pieces of data end up falling within the scope of indirect discrimination. Since the indirect discrimination doctrine was developed in order to recognise the specific effects that structural discrimination has on disadvantaged groups, it focuses on the group dimension and it is therefore necessary for the decision to affect a significant number of members of the group.

Proving cases of algorithmic discrimination by using the tools and framework provided by indirect discrimination is undoubtedly much more effective than trying to apply the direct discrimination framework since there are less transparency requirements needed to prove

indirect discrimination and opening the black box will generally not be necessary.[763] One of the positive aspects of algorithmic decision-making is that, if the output data is obtained, it is possible to statistically establish the way in which members of the disadvantaged group are affected.[764] This will likely mean that algorithmic discrimination cases will be based on statistical proof at a much higher rate than traditional discrimination cases. In fact, whenever an algorithm cannot be explained, and the actual measure that negatively impacts members of a protected group is therefore not known, cases of discrimination will probably have to be based on statistical data.

As it was already indicated when analysing direct algorithmic discrimination, new techniques will probably have to be developed and protocols drawn in judicial systems in order to prove algorithmic discrimination. One of said possible techniques is making sufficient information regarding the system available to plaintiffs alongside the system itself, so they can tinker and experiment with it.[765] To the extent that the training and test data, as well as other elements, are provided, it will be, in principle, possible for experts have enough information to prove the existence or lack of algorithmic discriminatory impact.[766]

3.2.2. Choosing how to measure indirect algorithmic discrimination

The section focused on "detecting and proving direct algorithmic discrimination" indicated that, said instances of discrimination can be measured as a function of individual fairness. However, individual fairness falls short in protecting members of disadvantaged groups from instances of indirect discrimination. The individual fairness metric does not consider the fact that members of disadvantaged groups share certain common habits, traits and attributes that are, in principle, non-sensitive features for they do not directly reflect protected group membership. These apparently non-sensitive features may be considered relevant towards predicting the target variable. This may result in some of the common attributes shared by many members of a protected group being coded as negative elements when making the relevant decision or altogether ignored so that only those specific attributes that are more

[763] XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *cit*., 2020, p. 171.
[764] *Ibidem*.
[765] KLEINBERG, J. *et al.*, "Discrimination in the age of algorithms", *cit.*, 2018, pp. 150-151.
[766] *Ibidem*.

common in members of dominant groups are coded as positive elements in the algorithm.[767] Since correlations between protected group membership and other features are not taken into account, these apparently non-sensitive features will show individuals to be differently situated (not similar), and therefore justify differences in treatment. Thus, in these cases, indirect discrimination will be considered justified or sometimes even go undetected.

Indirect discrimination must therefore be measured as a function of group parity. Group parity can either be expressed as statistical parity, also known as demographic parity or as accuracy parity, depending on the context in which the decision is made.

3.2.2.1. Statistical or demographic parity

Group fairness can be approached from the perspective of restrictive demographic or statistical parity, that is, the algorithm's outcome must more or less represent every group proportionally, according to their presence in society. Less restrictive forms of statistical parity can include establishing representation brackets that ensure that at least a certain percentage of protected groups is represented in automated decision-making.[768] This definition of group fairness ensures that protected group membership is not redundantly encoded through proxy variables for it focuses on equality of results.

This metric can be used in order to establish a *prima facie* case of indirect discrimination by indicating the percentage of members of a disadvantaged group that are negatively affected by the decision or by showing what fraction of the individuals negatively affected by the decision are members of a protected group. In the first case it would for example mean indicating that out of the total population of non-white people, 70% were negatively affected by the measure while only 20% of white people suffered the same negative consequences. The second possibility entails determining that out of all the people that have been negatively affected by the apparently neutral criterion, 75% were non-white and 25% were white.

The CJEU has repeatedly stated that the difference must be significant, as the measure must negatively affect considerably or far more members of the protected group than non-

[767] LOFTUS, J. *et al.*, "Causal reasoning for algorithmic fairness", 2018, p. 4. Available on 30th June 2020 at: https://arxiv.org/: "many variables vary along with protected attributes such as race or gender, making it challenging to find a distance measure that will not allow some implicit discrimination."
[768] BENT, J. R., "Is algorithmic affirmative action legal?", *The Georgetown Law Journal*, vol. 108, 2020, pp. 817-818.

members.[769] In this vein, AG Léger considered that proving that 60% of the individuals negatively affected by the measure were members of the protected group was insufficient to determine that a measure was discriminatory.[770] In general, the CJEU has established an apparently neutral measure to be indirectly discriminatory when members of the protected group constituted around 80 to 90% of the individuals negatively affected by the measure.[771]

There are many elements that courts can consider when determining whether a case can be considered discriminatory. For example, in *D.H. and others v. the Czech Republic,* the ECHR also took into consideration the percentage of Roma population in the Czech Republic to conclude that the representation of Roma children in special needs schools was very high if compared with the percentage of Roma children in all of the Czech Republic. The plaintiffs in this case argued that Roma children were being placed in schools for children with special needs at a much higher rate than non-Roma children and determined that, on average, pupils of Roma origin represented around 50 to 56% of the student population in special needs schools and that number went up to 90% in some schools, while Roma children in schooling age only represented around 2% of the country's student population.[772] This measure, which was taken on the basis on apparently neutral tests, was therefore deemed discriminatory by the ECHR as it impacted Roma children in a disproportionate manner.

Automated decision-making can enhance basing indirect discrimination claims on statistics as long as all output data is available. However, algorithmic opacity and complexity can also heavily hinder the possibility of assessing and analysing the apparently neutral provision, criterion or practice and the parameters upon which it is built. While statistics can be very useful in order to prove that an apparently neutral measure causes more harms to members of disadvantaged groups, Courts do not always require statistical proof and have sometimes determined the existence of a *prima facie* case of discrimination only by analysing the apparently neutral provision, criterion or practice.[773]

---

[769] CJEU Judgments 18th March 2014, C-167/12, C.D. v. S.T., paragraph 48 and 14th April 2015, C-527/13, Lourdes Cachaldora Fernández v Instituto Nacional de la Seguridad Social, paragraph 28.
[770] Opinion of Advocate General Léger delivered on 31st May 1995 on Case C-317/93 Inge Nolte v. Landesversicherungsanstalt Hannover, paragraph 58.
[771] WACHTER, S., "Affinity profiling and discrimination by association in online behavioural advertising", *cit.*, 2020 (forthcoming), p. 45.
[772] ECHR Judgment 13th November 2007, 57325/00, D.H. and others v. the Czech Republic, paragraphs 18 and 66.
[773] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", *cit.*, 2018, pp. 246-247.

Hence, if statistics are the only way in which algorithmic discrimination in opaque systems can be proven, it is crucial for courts to consider the possibility of lowering the threshold from which an algorithm is considered to be indirectly discriminatory. Furthermore, it is also very important to consider that, whereas in traditional cases of discrimination it might have been harder to exactly determine and measure how the member/non-member distribution of the apparently neutral provision's effects, in algorithmic discrimination cases it is possible to draw the exact data for the way in which the system impacts especially protected groups. In this context, it is necessary to refer to the CJEU's *Seymour-Smith* judgment in which the Court established that a *prima facie* case of indirect discrimination could be established if "statistical evidence revealed a lesser but persistent and relatively constant disparity over a long period between men and women..."[774]

Hence, taking into consideration the possibility of measuring outcomes in an exact and constant manner that algorithms offer and the CJEU's Judgment in the *Seymour-Smith* case, the threshold for considering indirect discrimination to be statistically proven could be lowered. If an algorithm is proven to systematically disadvantage members of a protected group at, for example, a 35% to 65% rate relative to non-members, the discriminatory nature of said system is undeniable and courts should therefore deem the system to be unlawful and contrary to the anti-discrimination legal framework.

3.2.2.2. When group parity does not work

However, statistical or demographic parity can generate very unfair results at the individual level and result in outcomes that are contrary to public interests, especially in its most restrictive versions. For example, taken to an extreme, statistical parity could justify determining that a criminal justice and prison system is discriminatory if an equally representative amount of members of all ethnicities and of each sex is not incarcerated. [775]

Demographic parity, as a criterion for group fairness can also enhance gaming behaviour if not properly controlled.[776] For example, a real estate firm may not want to rent or sell houses to members of ethnic minorities yet may have the obligation of showing its ads to an equally

[774] CJEU Judgment 9th February 1999, C-167/97, Regina v. Secretary of state for Employment, *ex parte* Nicole Seymour-Smith and Laura Perez, paragraph 60.
[775] BERK, R. *et al.*, "Fairness in criminal justice risk assessments: the state of the art", *Sociological Methods and Research*, 2018, p. 14.
[776] KROLL, J. A. *et al.*, "Accountable algorithms", *cit*., 2017, p. 686.

representative proportion of members of each ethnic group. In order to advance its own interests while complying with statistical parity mandates, the firm (or rather, the software developers it hires) may give instructions to the algorithm to maximise the chances that the ad will be shown to interested white individuals while minimising the chances that the ad will be shown to interested individuals belonging to ethnic minorities. Hence, while the ad will be shown to an equal proportion of members of each ethnic group, the automated system will ensure that members of ethnic minorities who are shown the ad are interested in its content.

Another issue that arises with regard to group fairness and, especially with statistical parity is that, when algorithmic discrimination is accurate, group fairness can lead to a significant reduction in accuracy. However, it is obviously necessary to keep in mind that even when discrimination is "accurate", it is generally still the result of the systemic discrimination that still endures in society. For instance, while black individuals do reoffend at a higher rate than white individuals in the US, the systemic discrimination that is embedded into US law enforcement structures necessarily brings about the question of what mechanisms can and should be adopted in order to redress the harms caused by existing structures of disadvantage and oppression against certain groups. As it will be argued later on, the notion of structural discrimination should be introduced as an interpretative criterion to draw policy and judge discrimination cases.

3.2.2.3. Accuracy parity

For contexts in which demographic parity criteria cannot be used or are not advisable, there are other options which focus on comparing the degree of accuracy of algorithms across different groups by carrying out ongoing analyses that compare inter and intragroup false positives and negatives.[777] European courts should accept accuracy parity as way to measure indirect discrimination. These measures of fairness aim to determine equality through accuracy in the predictions made for disadvantaged groups in comparison to non-disadvantaged groups. References to recidivism risk algorithms will be used in order to explain the different versions of accuracy parity but said examples could be exchanged for a number of target variables such as job performance or probability of loan default.

The following table represents the possible situations that can result out of the predictions made by an algorithm designed to predict recidivism. The table below is merely mean to be a

[777] BERK, R. *et al.*, "Fairness in criminal justice risk assessments…", *cit.* 2018, pp. 14-15.

visual aid to understand the different ways in which accuracy parity can be measured which are explained below:

	Predicted to re-offend	Predicted not to reoffend
Did re-offend	a	b
Did not re-offend	c	d

i) *True positive and negative rate parity*

A way in which it is possible to measure this version of group fairness is by comparing whether the true positive rate is the same for different groups (for each group, out of the total number of individuals who actually reoffended, what percentage had been predicted to reoffend? [a/a+b]) and whether the rate of false negatives is the same for different groups (for each group, out of the total number of individuals who did not reoffend, what percentage had been predicted not to reoffend? [d/c+d]).[778] This is what BERK *et al.* label "conditional procedure accuracy equality"[779] and HARDT *et al.* define as "equality of opportunity".[780]

ii) *Rate of accuracy in the prediction of positives and negatives*

A second possibility is to measure and compare the rate of accuracy in the prediction of positives across groups (for each group, out of the people predicted to reoffend, what percentage did reoffend? [a/a+c]) and the rate of accuracy in the prediction of negatives across groups (for each group, out of the people predicted not to reoffend, what percentage did not reoffend? [d/b+d]).[781] This is what BERK *et al.* label "conditional use accuracy equality".[782]

[778] BENT, J. R., "Is algorithmic affirmative action legal?", *cit.*, 2020, p. 818.
[779] BERK, R. *et al.*, "Fairness in criminal justice risk assessments…", *cit.* 2018, p. 14.
[780] HARDT, M., PRICE, E. & SREBRO, N., "Equality of opportunity in supervised learning", paper presented at the *30th Conference on Neural Information Processing Systems*, Barcelona, 2016, p. 2. Available on 2nd July 2020 at: http://papers.nips.cc/
[781] BENT, J. R., "Is algorithmic affirmative action legal?", *cit.*, 2020, p. 818.
[782] BERK, R. *et al.*, "Fairness in criminal justice risk assessments…", *cit.* 2018, p. 14.

iii) *False positive to false negative ratio and false negative to false positive ratio*

The third viable version of group parity can be measured through the ratio of false positives to false negatives and of false negatives to false positives for each group.[783] Using this metric can help identify for which group it is costlier (difficult) to get a false negative (a false negative is the best possible outcome for those being subjected to the predictive system). For instance, if for every 10 false positives there are 2 false negatives for black men, but for every 10 false positives there are 4 false negatives for white men. If false negatives are obviously costlier for protected groups, using this measure may be useful to argue for modifying either the input data or the system itself to reduce the false positive rate or the rate of error in the prediction of positives for members of protected groups.

iv) *Combined rate of true positives and negatives: overall accuracy*

There are other possibilities, such as calculating and comparing "overall accuracy equality"[784] in each group, which is done by calculating the percentage of true positives and negatives as a fraction of the total population of each group examined. For instance, out of all the white individuals whose recidivism risk was predicted, what was the percentage of predictions that were wrong? The problem with this measure is that it does not differentiate between true positives and true negatives and thus treats them as equally desirable. However, establishing differences between the desirability of true positives and true negatives (or the undesirability of false positives and false negatives) is very useful, if not essential, in order to translate policy preferences in the calibration of algorithms.

3.2.2.4. Is total group fairness possible?

Total group fairness is represented by a function in which the mandates set by statistical parity and all forms of accuracy parity are met. Total group fairness has been proven to be impossible except in very artificial contexts.[785] Consequently, specific policy decisions will

[783] BENT, J. R., "Is algorithmic affirmative action legal?", *cit.*, 2020, p. 818; BERK, R. *et al.*, "Fairness in criminal justice risk assessments…", *cit.* 2018, p. 15.
[784] BERK, R. *et al.*, "Fairness in criminal justice risk assessments…", *cit.* 2018, p. 14.
[785] BERK, R. *et al.*, "Fairness in criminal justice risk assessments…", *cit.* 2018, p. 15; Chouldechova, A., "Fair prediction with disparate impact: a study of bias in recidivism prediction instruments", 2016. Available on 9th April 2019 at: https://arxiv.org/; CORBETT-DAVIES, S., PIERSON, E. & GOEL, S., "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear", *The Washington Post*, 17th October 2015. Available on 9th April 2019 at: https://www.washingtonpost.com/; CORBETT-DAVIES, S. & GOEL, S., "The measure and mismeasure of fairness: a critical review of fair machine learning", 2018, pp. 1-25. Available on 6th May 2020 at: https://arxiv.org/; FRIEDLER, S. A. *et al.*, "On the (im)possibility of

have to made in order to determine what constitutes the best measure of parity depending on the context

3.3. INDIRECT DISCRIMINATION BY ASSOCIATION

As with direct discrimination, the CJEU's *CHEZ*[786] judgment established that indirect discrimination by association is also prohibited. In the *CHEZ* case, the plaintiff lived in an area that was mostly populated by members of the Roma community but she did not belong to said community herself. What happened in said case is that the electric company had placed the meters in her area out of reach because of prejudices held against Roma people and the plaintiff could therefore not access her electricity meter. Including this case under the special protection offered by anti-discrimination legislation is particularly relevant as it was already indicated with regard to the *Coleman* case, for it widens the personal scope of application of the right to equality and non-discrimination of members of disadvantaged groups, extending it to those who are not members of such groups but suffer discrimination by perception or association.[787]

It is also worth mentioning that, in this case, the CJEU left open the possibility of interpreting whether this constituted a case of direct or indirect discrimination due to the fact that the contentious practice, while it affected both members and non-members of the Roma community, was adopted with the specific aim of preventing tampering by the Roma population. It is therefore possible to argue that the specially protected category determined the adoption of the discriminatory practice.

In any case, as it was already indicated when addressing the *Coleman* ruling, this particular judgment opens up the door to recognising the discriminatory nature of protected group affinity detection tools used in targeted advertising, such as the ones that were used by Facebook.[788] However, it is important to highlight the difficulties that might be encountered

---

fairness", 2016. Available on 15th June 2020 at: https://arxiv.org/; KLEINBERG, J., MULLAINATHAN, S., & RAGHAVAN, M., "Inherent trade-offs in the fair determination of risk scores", 2016. Available on 18th February 2019 at: https://arxiv.org/

[786] CJEU Judgment 16th July 2015, C-83/14, CHEZ Razpredelenie Bulgaria AD v. Komisia za zashtita ot diskriminatsia.

[787] WACHTER, S., "Affinity profiling and discrimination by association in online behavioural advertising", *cit.*, 2020 (forthcoming), pp. 36-41.

[788] ANGWIN, J. & PARRIS Jr., T., "Facebook lets advertisers exclude users by race", *cit.*, 2016; ANGWIN, J., TOBIN, A. & VARNER, M., "Facebook (still) letting housing advertisers exclude users by race", *cit.*, 2017; GILLUM, J. & TOBIN, A., "Facebook won't let employers, landlords or lenders discriminate in ads anymore", *cit.*, 2019.

when proving indirect algorithmic discrimination by association due to the fact that plaintiffs will have to prove that the apparently neutral practice causes more harm to actual and erroneously identified members of the protected group than to non-members.[789] In some cases, it will not be possible to access said information without opening the black box for it will be very hard to determine which non-members have been perceived to be members of the group.

3.4. JUSTIFICATIONS TO INDIRECT ALGORITHMIC DISCRIMINATION AND THE PROBLEM WITH ACCURATE DISCRIMINATION

The main difference between the legal treatment of direct and indirect discrimination is that the latter allows for a greater margin of justification, especially within the framework of the CJEU. Hence, once a *prima facie* case of indirect discrimination has been established it is possible for the respondent to justify the discriminatory measure. Both the ECHR and CJEU admit some form of objective justification to indirect discrimination.[790]

In order for a justification to be considered sufficient, the apparently neutral provision, criterion or practice must serve a legitimate aim and must also pass the three-pronged proportionality test in which the measure must be proven i) to be appropriate (suitable) to accomplish its aim; ii) to be necessary, that is, there must be no other mechanisms which would serve the same purpose while not affecting the right to equality and non-discrimination and, iii) it also has to pass a *strictu sensu* proportionality test, whereby the different interests at play are balanced.[791]

The EU Equality Directives only mention the two first parts of the three-pronged proportionality test.[792] This has led the CJEU to carry out the second and third parts of the proportionality test in a joint manner in many cases and, sometimes, even to altogether ignore the third part when it has considered it was sufficient to only carry out the two first parts of the test.[793] However, since it established the three-step test in the *Bilka-Kaufhaus* ruling,[794]

[789] WACHTER, S., "Affinity profiling and discrimination by association in online behavioural advertising", *cit.*, 2020 (forthcoming), p. 41.
[790] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", *cit.*, 2018, p. 94.
[791] CRAIG, P. & DE BURCA, G., *EU Law: Text Cases and Materials*, New York, Oxford University Press, 5th ed., 2011, p. 526.
[792] See article 2 in each of the EU Equality Directives (Racial Equality Directive, Employment Equality Directive, Gender Employment Equality Directive and Gender Goods and Services Directive).
[793] CRAIG, P. & DE BURCA, G., *EU Law…*, *cit*., 2011, p. 526.

the Court has referred to the need to carry out the "balancing of interests" test in order to consider a discriminatory measure sufficiently justified in several cases.[795] Moreover, considering the relevance that the three-pronged version of the proportionality test has in member state law[796] and given the importance that balancing interests has with regard to algorithmic discrimination, the three steps or prongs will be analysed separately.

3.4.1. Legitimate aim

The CJEU has not established a consistent standard of what a legitimate aim constitutes but has rather subjected said determination to considering that the measures are appropriate, necessary and proportional.[797] In fact, given the weight that the Court places on the other requirements for justification, it has generally been quite lenient in determining discriminatory measures to serve legitimate aims.[798] For instance, in the *Bilka-Kaufhaus* judgment, the Court opened the door to the possibility of justifying a measure which excluded part-time workers from its pension scheme in order to incentivise full-time work even though this measure negatively impacted many more women than men.[799] In the same vein, the CJEU's ruling in the *Achbita* case established that "the pursuit by the employer, in its relations with its customers, of a policy of political, philosophical and religious neutrality",[800] constitutes a legitimate aim as it "relates to the freedom to conduct a business that is recognised in Article 16 of the Charter".[801] Hence, overall, the CJEU has considered economic interests as valid legitimate aims within the private sector.

---

[794] CJEU Judgment 13th May 1986, C-170/84, Bilka – Kaufhaus GmbH v. Karin Weber von Hartz, *cit.* paragraphs 37 and 38.
[795] CJEU Judgment 14th March 2017, C-157/15, Samira Achbita and Centrum voor gelijkheid van kansen en voor racismebestrijding v G4S Secure Solutions NV, paragraphs 35 and 43; CJEU Judgment 22nd November 2005, C-144/04, Werner Mangold v. Rüdiger Helm, paragraph 65.
[796] NEWTON, M. & MAY, L. A., *Proportionality in International Law*, Oxford, Oxford University Press, 2014, p. 156; CRAIG, P. & DE BURCA, G., *EU Law…*, *cit*., 2011, p. 526.
[797] CJEU Judgment 15th May 1986, C-222/84, Johnston v. Chief Constable of the Royal Ulster Constabulary; CJEU Judgment 17th July 2014, C-173/13, Maurice Leone, Blandine Leone v. Garde des Sceaux, ministre de la Justice, Caisse nationale de retraite des agents des collectivités locales.
[798] TOBLER, C., *Limits and potential of the concept of indirect discrimination*, Luxembourg, Publications Office of the EU, 2008, p. 43.
[799] CJEU Judgment 13th May 1986, C-170/84, Bilka – Kaufhaus GmbH v. Karin Weber von Hartz, paragraph 37: "A department store company may justify the adoption of a pay policy excluding part-time workers, irrespective of their sex, from its occupational pension scheme on the ground that it seeks to employ as few part-time workers as possible, where it is found that the means chosen for achieving that objective correspond to a real need on the part of the undertaking, are appropriate with a view to achieving the objective in question and are necessary to that end."
[800] CJEU Judgment 14th March 2017, C-157/15, Samira Achbita and Centrum voor gelijkheid van kansen en voor racismebestrijding v G4S Secure Solutions NV, paragraph 37.
[801] *Idem*, paragraph 38.

With regard to the public sector, discriminatory measures have been ruled to respond to legitimate aims when they fell under the umbrella of public policy objectives such as "ensuring coherence of the tax system" or "guaranteeing a minimum replacement income"[802] and, especially, when they "reflect a legitimate social-policy objective".[803] Conversely, the Court has clearly stated that budgetary considerations cannot constitute legitimate aims for a discriminatory policy unless they are supported by another public policy objective.[804]

In theory, as long as the aim itself is not discriminatory, or does not mask discriminatory intent in an obvious manner, it will be deemed legitimate. The aims for which automated systems are used, which include determining creditworthiness, the best candidates in recruitment processes or matching individuals with products they might be interested in will therefore logically be considered to be legitimate and pass the first part of the justification test.[805] However, for instance using algorithms with the objective of carrying out predatory advertising practices, that is, targeting members of socially vulnerable groups in order to sell them low quality products and services, could be ruled to not comply with the legitimate aim requirement set out in order to justify discrimination. This is, of course, as long as the case falls within the scope of application of EU law.

It is nonetheless necessary to point out whether the use of automated decision-making with more general aims, such as "profit improvement", should be deemed legitimate or not. This question becomes particularly relevant when it comes to the use of algorithmic price discrimination.[806] The possibility of setting different prices according to what each person is willing to pay is enhanced through the use of machine learning algorithms. The use of differential pricing is generally expected to have redistributive effects.[807] However, these methods are sometimes used in an abusive manner, damaging already vulnerable groups,

[802] MALISZEWSKA-NIENARTOWICZ, J., "Direct and indirect discrimination in European Union law – how to draw a dividing line?", *International Journal of Social Sciences*, vol. 3, No. 1, 2014, pp. 44-45.
[803] CJEU Judgment 17th July 2014, C-173/13, Maurice Leone, Blandine Leone v. Garde des Sceaux, ministre de la Justice, Caisse nationale de retraite des agents des collectivités locales, paragraph 53.
[804] CJEU Judgment 11th November 2014, C-530/13, Leopold Schmitzer v. Bundesministerin für Inneres, paragraph 41 and CJEU Judgment 21st July 2011, Joined Cases C-159/10 and C-160/10, Gerhard Fuchs (C-159/10), Peter Köhler (C-160/10) v. Land Hessen, paragraph 74: "…while budgetary considerations can underpin the chosen social policy of a member state and influence the nature or extent of the measures that the member state wishes to adopt, such considerations cannot in themselves constitute a legitimate aim within the meaning of Article 6(1) of Directive 2000/78."
[805] HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, p. 1161.
[806] ZUIDERVEEN BORGESIUS, F., "Price discrimination, algorithmic decision-making, and european non-discrimination law", *European Business Law Review* (Forthcoming), 2020, pp. 1-29. Available on 10th June 2020 at: https://ssrn.com/
[807] ELEGIDO, J. M., "The ethics of price discrimination", *Business Ethics Quarterly*, vol. 21, No. 4, 2011, p. 639.

thereby further deepening already existing situations of inequality and discrimination. For example, VALENTINO-DEVRIES, SINGER-VINE and SOLTANI found that an algorithm which varied prices according to where it believed the customer was located was leading to people who lived in higher-income areas to receive larger discounts than people who lived in low-income areas.[808]

It is at this point relevant to recall the *Achbita* judgment,[809] which considered offering an appearance of religious neutrality a legitimate aim under the undertaking's freedom to conduct a business, even though the legitimate aim itself is discriminatory, for under said veneer of neutrality there is a clear disregard for the particular experience of Muslim women. It is not possible to ignore the fact that most discussions regarding religious dress codes carried out in Europe throughout the past decades have focused on the Islamic headscarf, which places the women of this very specific religion at a clear disadvantage.[810]

Hence, if this very clearly discriminatory aim was deemed legitimate by the Court on the basis of the defendant's freedom to conduct a business, it is also likely that the use of price discrimination in order to improve profit is also considered a legitimate aim.[811]

3.4.2. The first prong of the proportionality test: appropriateness or suitability

In order to justify a discriminatory algorithm, it will be necessary to prove it is appropriate to achieve the objective at hand, that is, that the automated system is suitable to predict whatever the target variable (legitimate aim) is in the specific case. There are two possible variants to the appropriateness test when applied to algorithmic discrimination. Courts could either decide that the whole system constitutes the apparently neutral provision, criterion or practice or could choose to establish a more accurate appropriateness test by assessing the specific variables causing the discriminatory results.[812]

If Courts choose the second option, respondents will be required to provide evidence that the variables which determine the discriminatory outcome have predictive value with regard to

[808] VALENTINO-DEVRIES, J., SIGER-VINE, J., SOLTANI, A., "Websites vary prices, deals based on users' information", *The Wall Street Journal*, 24th December 2012. Available on 12th February 2019 at: https://www.wsj.com/
[809] CJEU Judgment 14th March 2017, C-157/15, Samira Achbita and Centrum voor gelijkheid van kansen en voor racismebestrijding v G4S Secure Solutions NV.
[810] CUYPERS, D., "Religion, discrimination, the head scarf and labour law", *ERA Forum*, vol. 19, 2019, p. 436.
[811] ZUIDERVEEN BORGESIUS, F., "Price discrimination, algorithmic decision-making, and european non-discrimination law", *cit.*, 2020, pp. 17-19.
[812] HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, p. 1161.

the target variable.[813] If the variables leading to discriminatory results are not proven to have predictive value in accomplishing the legitimate aim, the algorithm will fall under the scope of direct discrimination by inference, as it was previously discussed.[814] However, if the variables are proven to have predictive value, they will pass the appropriateness test for they will have been proven to be suitable in achieving the legitimate aim.

Nevertheless, there are a series of elements that make it more likely that courts will consider the system to constitute the apparently neutral practice.[815] Firstly, algorithmic systems can be highly complex and their discriminatory outcomes are in many cases the result of combining a very large number of variables. Secondly, sometimes, the discriminatory output may not result from the way in which the variables are weighed and defined but from other elements, such as biases contained in the dataset. Thirdly, algorithms will generally be presented as single items by their developers.

In these cases, if the algorithm as a whole is considered and analysed as the apparently neutral practice, the appropriateness or suitability of the automated system will be fairly easy to prove for respondents. The CJEU has in the past required specific evidence to prove the suitability of an indirectly discriminatory measure in accomplishing the legitimate objective and has determined "mere generalisations concerning the capacity of a specific measure to encourage" [816] the legitimate aim to be insufficient to pass the appropriateness test.[817]

Algorithms will easily pass the suitability or appropriateness test, as it will only be necessary to demonstrate the algorithm's accuracy, especially in simpler systems, by validating it through calculating different relationships between the percentage of false/true positives and false/true negatives.[818] In complex systems, such as recidivism risk algorithms, it will be necessary to measure the combinations, types and costs of errors which will render measuring

[813] GRIMMELMANN, J. y WESTREICH, D., "Incomprehensible discrimination", *cit.*, 2017, p. 176.
[814] See section 2.2 in this Chapter: "What constitutes direct algorithmic discrimination?"; GRIMMELMANN, J. y WESTREICH, D., "Incomprehensible discrimination", *cit.*, 2017, p. 176.
[815] HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, pp. 1161-1162.
[816] CJEU Judgment 9th February 1999, C-167/97, Regina v. Secretary of state for Employment, *ex parte* Nicole Seymour-Smith and Laura Perez, paragraph 76: "Mere generalisations concerning the capacity of a specific measure to encourage recruitment are not enough to show that the aim of the disputed rule is unrelated to any discrimination based on sex nor to provide evidence on the basis of which it could reasonably be considered that the means chosen were suitable for achieving that aim".
[817] FREDMAN, S., *Discrimination Law*, *cit*., 2011, p. 194.
[818] SCATAMBURLO, T., CHARLESWORTH, A. & CRISTIANINI, N., "Machine decisions and human consequences", *cit.*, 2019, pp. 62-68.

accuracy slightly more difficult, but nonetheless feasible.[819] It is important to highlight that, at this point, we would not be measuring accuracy parity but the system's overall accuracy.

Courts should however establish an additional requirement that the accuracy test is carried out against real world data and not deem sufficient the results yielded from testing the algorithm against the test dataset, which will be more likely to incorporate the biases in previous decision-making processes as well as the biases held by the algorithm designers.[820] Otherwise, there is a significant risk that indirectly discriminatory algorithmic systems that are not suitable to accomplish the legitimate aim (not predictive of the target variable) pass the appropriateness test because they are tested against biased datasets. Testing the algorithm against real world data does not mean that said risk will be completely eliminated, but it will certainly provide a larger and more accurate suitability evidence base.

Courts and policymakers should establish a minimum level of accuracy that algorithms should comply with and decide what types of errors should be prioritised depending on the objective for which the system is being used. These decisions are highly relevant. For instance, if an employment performance assessment system classifies workers as high-performers and low-performers, when the algorithm is tested for accuracy, two types of errors will be defined, high-performers who had been predicted to be low-performers (error 1), and low-performers who had been predicted to be high-performers (error 2). Those in charge of designing the evaluation system will generally place a different value on each type of error. If error 1 is labelled as less harmful than error 2, the systems results for accuracy measurements will constantly undervalue the persistence of high-performer misclassification which will also lead to the perpetuation of group-disadvantage if the algorithm is biased against the members of protected groups. It is therefore essential to introduce accuracy parity as part of the appropriateness test and evaluate what types of errors hurt disadvantaged groups the most.

3.4.3. The second prong of the proportionality test: necessity

In the *Bilka-Kaufhaus* judgment, the CJEU established that in order for the indirectly discriminatory practice to pass the proportionality test, it must show that "the means chosen for achieving [the legitimate] objective correspond to a real need on the part of the

[819] *Ibid.*, p. 68.
[820] HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, pp. 1161-1162.

undertaking." [821] An indirectly discriminatory measure is deemed necessary when there is no alternative, less discriminatory measure, to achieve the same aim.[822]

The burden of proving that a less discriminatory measure does not exist lies on the respondent,[823] which should lead courts to require defendants to provide evidence that they did consider other alternatives to the algorithmic system that was finally employed. It is essential to establish procedural rules that prevent courts from considering that defendants pass the necessity test by simply arguing that, since machine learning algorithms are generally more accurate than traditional forms of decision-making, there is no other less discriminatory and equally accurate way in which to achieve the same aims.[824]

The increasing concern with algorithmic discrimination has brought along a field of study focused on developing non-discriminatory algorithms and algorithms that can help detect and correct biases in automated systems.[825] The availability of this type of technology means that respondents should at least prove that they tried other automated systems or other mechanisms aimed towards preventing discriminatory results.

Additionally, proof that the dataset was not biased should also be presented. The large amounts of information that are currently available do not necessarily entail that all datasets are equally accurate. Datasets that have undergone lower levels of scrutiny and review are sold at cheaper prices than datasets that are examined in greater detail and that are only put on the market when they meet very high standards and are therefore more accurate. Consequently, it is likely that organisations choose less accurate and cheaper datasets which, overall, when processed, do achieve the aims the automated process is set out to attain.[826]

---

[821] CJEU Judgment 13th May 1986, C-170/84, Bilka – Kaufhaus GmbH v. Karin Weber von Hartz, paragraph 37.

[822] CJEU Judgment 16th July 2015, C-83/14, CHEZ Razpredelenie Bulgaria AD v. Komisia za zashtita ot diskriminatsia: "…such a measure would be capable of being objectively justified … only if that measure did not go beyond what is appropriate and necessary to achieve those legitimate aims and the disadvantages caused were not disproportionate to the objectives thereby pursued. That is not so if it is found … that other appropriate and less restrictive means enabling those aims to be achieved exist…"

[823] FREDMAN, S., *Discrimination Law*, *cit*., 2011, p. 194.

[824] HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, p. 1162.

[825] CORBETT-DAVIES, S. & GOEL, S., "The measure and mismeasure of fairness…", *cit.*, 2018, pp. 1-25. DWORK, C. *et al.*, "Fairness through awareness", *cit.*, 2012, pp. 214-226; FELDMAN, M. *et al.*, "Certifying and removing disparate impact", *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 259-268; ŽLIOBAITĖ, I. & CUSTERS, B., "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models", *Artificial Intelligence & Law*, vol. 24, No. 2, 2016, pp. 183-201.

[826] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 689.

Tests should therefore be run in order to prove that the dataset used in the decision-making process is not biased.[827]

Finally, a comparative analysis with other similar private and public organisations should also be carried out in order to determine whether an alternative system that yields less discriminatory results is being used by any of those organisations in order to achieve the same purpose.[828]

Courts must not be lenient in allowing defendants to contend that there is not a less discriminatory alternative: the rapid pace at which technological development is taking place means that if we accept that very accurate profiles of individuals are being carried out, providing firms with significant profit gains they must also, in exchange develop the technology necessary to guarantee that fundamental rights are respected.

3.4.3.1. Establishing necessity through the "best available techniques" criterion[829]

An interesting approach to the problems generated by algorithms and their possible solutions is the one that draws parallelisms between environmental pollution and the issues generated by data processing technologies.[830] For instance, BEN-SHAHAR, argues that, like environmental pollution, what he labels as "data pollution" produces public harms[831] that cannot properly be addressed by private law.[832] The existence of some similarities between both phenomena are further reinforced by the fact that the European data protection regulatory framework, which will be addressed in the dissertation's second part, uses some of the regulatory tools, such as impact assessments or information rights, that are typically employed in the regulation of environmental pollution.[833]

Applying the concept and mandate of "best available techniques", developed in environmental law, can prove extremely useful when carrying out the second prong of the

[827] HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, pp. 1162-1163.
[828] CJEU Judgment 16th July 2015, C-83/14, CHEZ Razpredelenie Bulgaria AD v. Komisia za zashtita ot diskriminatsia, paragraph 121: "…the KZD submitted in its observations that other electricity distribution companies have given up the practice at issue, giving preference to other techniques for the purpose of combating damage and tampering…"
[829] I would like to thank professor Gabriel Doménech for this idea.
[830] BEN-SHAHAR, O., "Data pollution", *Journal of Legal Analysis*, vol. 11, 2019, pp. 104-159.
[831] *Idem*, pp. 110-118.
[832] *Idem*, pp. 118-131.
[833] BINNS, R., "Data protection impact assessments: a meta-regulatory approach", *International Data Privacy Law*, vol. 7, No. 1, 2017, p. 23; MAZUR, J., "Automated decision-making and the precautionary principle in EU law", *Baltic Journal of European Studies Tallinn University of Technology*, vol. 9, No. 4, 2019, pp. 3-18.

proportionality test in order to determine whether the use of an algorithm that produces discriminatory results is necessary.

The "best available techniques" rule was generalised in the EU under the 1996 Directive on integrated pollution prevention and control[834] and is currently contained in the 2010 Directive on industrial emissions,[835] which defines "best available techniques" as follows:[836]

> "…'best available techniques' means the most effective and advanced stage in the development of activities and their methods of operation which indicates the practical suitability of particular techniques for providing the basis for emission limit values and other permit conditions designed to prevent and, where that is not practicable, to reduce emissions and the impact on the environment as a whole:
>
> (a) 'techniques' includes both the technology used and the way in which the installation is designed, built, maintained, operated and decommissioned;
> (b) 'available techniques' means those developed on a scale which allows implementation in the relevant industrial sector, under economically and technically viable conditions, taking into consideration the costs and advantages, whether or not the techniques are used or produced inside the member state in question, as long as they are reasonably accessible to the operator;
> (c) 'best' means most effective in achieving a high general level of protection of the environment as a whole…"

Hence, the "best available techniques" rule does not aim to force firms to use the last state of the technology as it acknowledges the unreasonable costs that said mandate would have. Said concept can be identified with the references made in the GDPR to "the state of the art" and "cost of implementation" of measures to safeguard individuals' right to data protection. For instance, article 25 of the GDPR establishes that:

> "Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the

[834] Directive 96/61/EC of 24 September 1996 concerning integrated pollution prevention and control.
[835] Directive 2010/75/EU of 24 November 2010 on industrial emissions (integrated pollution prevention and control).
[836] Article 3.10 of the 2010 Directive on industrial emissions.

controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects".

Similarly, using non-discriminatory algorithms may not always be possible because said technology may be too costly for developers. However, by establishing a series of common standards and shared knowledge of what the "best available techniques" are in the development of non-discriminatory algorithms used for different purposes and in different sectors, it should be possible to set a series of minimum requirements with regard to compliance with the rights to equality and non-discrimination. More importantly, by continuously updating the reference documents in which the "best available techniques" are indicated, regulators and Courts will have the closest thing to an objective indicator of what are the technologies that should be used in order to consider that there is no alternative, less discriminatory measure, to achieve the same aim.

The "best available techniques" mechanism may also be used in order to ensure that algorithms respect other fundamental rights and public interests. The way in which said mechanism should be developed will be addressed at the end of the dissertation's second part.

3.4.4. The third prong of the proportionality test: *strictu sensu* proportionality and accurate indirect algorithmic discrimination

The last part of the proportionality test requires balancing the competing interests at play.[837] In general terms, these conflicts are generally presented as a trade-off between equality and efficiency or equality and freedom or a combination of both, for instance, when the right to non-discrimination competes with the freedom to conduct a business.

[837] Opinion of Advocate General Kokott delivered on 31st May 2016 on Case C-157/15 Samira Achbita and Centrum voor gelijkheid van kansen en voor racismebestrijding v. G4S Secure Solutions NV, paragraph 112: "Finally, the third issue to be examined is proportionality *sensu stricto*. According to that principle, measures must not, even if they are appropriate and necessary for achieving legitimate objectives, give rise to any disadvantages which are disproportionate to the objectives pursued. In other words, therefore, it must be ensured that a ban such as that at issue does not have the effect of unduly prejudicing the legitimate interests of employees. Ultimately, this means that a fair balance must be struck between the conflicting interests of employees such as Ms Achbita, on the one hand, and undertakings such as G4S, on the other."

In order for an algorithm to pass the two previous parts of the proportionality test it will be necessary for it to be accurate or, at least, to attain the aim for which the algorithm is deployed in a more accurate manner than the traditional or alternative mechanisms that can be used for the same purpose (within the reasonable demands set by the best available techniques model). In that case, even if the algorithm is discriminatory, given the fact that there is no equally accurate and less discriminatory decision-making system, it will be necessary for courts to evaluate, on a case-by-case basis, whether the interests of the respondent should be prioritised or if the algorithm should be modified in order to introduce some form of affirmative action or accommodation to ensure that the same real opportunities are offered to members and non-members of disadvantaged groups.

The accurate discrimination of protected groups is in many cases the result of the structural discrimination suffered by those individuals in combination with the way in which the instructions fed into the algorithm are framed.[838] It is important to take this into consideration because, even if an algorithm seems to accurately predict say, for example, that racial minorities will do worse in university, said reality is the result of a history of structural discrimination that has systematically deprived members of disadvantaged groups from having the same real opportunities as non-members.[839] Moreover, the actual inaccuracy of apparently accurate statistical discrimination in traditional forms of decision-making[840] will also very likely take place (and go undetected) in automated decision-making given the many possibilities that there are for introducing biases through both the development and deployment stages of automated systems.[841]

The EU has largely failed to acknowledge this reality, seeing as the Commission's guidelines issued after the *Test-Achats* case, established that "the use of risk factors which might be correlated with gender therefore remains possible, as long as they are true risk factors in their own right."[842] Conversely, following this interpretative line may also favour disadvantaged groups if it places the burden of proving that a factor correlated with sex or another protected

[838] FRIEDLER, S. A. *et al.*, "On the (im)possibility of fairness", *cit.*, 2016, p. 7.
[839] AÑÓN ROIG, M. J., "Principio antidiscriminatorio y determinación de la desventaja", *cit.*, 2013b, pp. 145-151.
[840] BOHREN, J. A. *et al.*, "Inaccurate statistical discrimination", *cit.*, 2019.
[841] WICK, M. *et al.*, "Unlocking fairness: a trade-off revisited", paper presented at the conference on *Advances in Neural Information Processing Systems (NIPS 2019)*, 2019. Available on 16th June 2020 at: https://papers.nips.cc/
[842] EU COMMISSION, "Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats)", Official Journal of the EU, 13th January 2012, paragraph 17.

category is, in fact, a risk factor. Moreover, these particular guidelines seem to be largely oriented to cases in which men pay, on average, higher insurance premiums than women which is not as problematic, seeing as social structures do not disadvantage men in a systematic manner to the same extent as they disadvantage women.

Accurate discrimination should be considered within the context and as the result of historical group disadvantage and structural discrimination. Moreover, it is practically impossible to establish, without any doubt, that the results obtained are really accurate for we cannot determine all possible ways to measure a specific social phenomenon. However, since the notion and framework of indirect discrimination is built without regard for the existence of a historical structure of subordination and courts do not generally introduce the structural discrimination perspective into their judgments,[843] it is in fact likely that instances of accurate discrimination are considered justified.

Amongst the criteria considered in this part of the proportionality test, courts should therefore include the "reasonability" of forcing respondents to use non-discriminatory models.[844] I propose that a cost-benefit analysis is carried out as part of the reasonability test in cases of (apparently) accurate algorithmic discrimination. Said analysis should be developed through the examination of three elements.

The first element evaluated should involve the costs of discriminatory algorithmic systems, that is, the risks for all fundamental rights and other public interests generated by the purpose for which the system is being used. For instance, algorithms used in the criminal justice system can cause significantly more damage to the fundamental rights of individuals than those used in targeted advertising.[845]

Secondly, in this part of the test it is also necessary for courts to consider the risks (costs) generated by subjecting certain types of decisions to automation. Humans, while generally less accurate than machine learning algorithms, are capable of considering the specific particularities of each case and observing details that automated systems may not detect. As a result of this ability of observing the intangible, even though humans are heavily biased, said biases impact singular decisions in a different manner. Contrariwise, if an algorithm is biased

[843] AÑÓN ROIG, M. J., "Principio antidiscriminatorio y determinación de la desventaja", *cit.*, 2013b, pp. 145-151.
[844] HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, p. 1164.
[845] *Ibidem.*

against the members of a disadvantaged group, it will systematically make decisions that perpetuate said disadvantage. It is therefore necessary that when Courts carry out the "balancing of interests" part of the proportionality test, they also consider the value of the human capacity in grasping certain intangible aspects that will probably not be processed by machines. In this context, it is relevant to bring up the study carried out by DE-ARTEAGA *et al.* shows that social workers were able to detect when an algorithm used to assist child welfare services made mistakes based on erroneous child risk estimates.[846] In this case, the "humans in the loop" had access to the children's files and all data regarding each particular case. This example is particularly important with regard to the increasing automation of welfare services.

Finally, courts should also consider the cost of not deploying the algorithm or reducing its accuracy or, to put it differently, the benefits of discrimination in each specific case analysed. Exceptions to the "reducing accuracy in favour of non-discrimination" rule would, for example, include algorithms designed to predict the probability that victims of violent crimes are newly attacked. For instance, if the algorithm used in Spain to predict that victims of gender-based violence will suffer new attacks by their aggressors,[847] on average predicts a higher level of risk for women whose attackers are immigrants. In this case there is a very specific risk for the fundamental right to life and physical and mental integrity. In addition, the victim also belongs to an especially protected group, which further justifies prioritising efficiency over equality.

Another case in which a certain level of discrimination would be justified is in cases in which the necessity of the measure is very straightforward, such as when socioeconomic status is considered when granting loans. Socioeconomic status is not currently comprehended as one of the specially protected characteristics included in the EU Equality Directives and is not even directly included under the European Convention of Human Rights and the Charter of Fundamental Rights of the EU. Without discussing whether this category should be explicitly included, which is something that will be addressed later on, lower socioeconomic status is also generally a proxy for other protected groups, such as racial minorities and women. Consequently, using socioeconomic status as a variable in determining loan eligibility, results

[846] DE-ARTEAGA, M. FLOGLIATO, R. & CHOULDECHOVA, A., "A case for humans-in-the-loop…", *cit.*, 2020, pp. 1-12.
[847] CABALLÉ-PÉREZ, M. *et al.*, "El quebrantamiento de las órdenes de protección en violencia de género: análisis de los indicadores de riesgo mediante el formulario vpr4.0", *cit.*, 2020, pp. 63-72.

in smaller loans with stricter conditions being granted to members of disadvantaged groups. However, evaluating individuals' financial situation is essential in order to determine loan eligibility. Nonetheless, it is vital to ensure that disadvantaged group membership has no relevance in determining the final decision and that the differences in outcomes between members and non-members are only those that can be justified by the differences in their financial situation. For these cases, the individual fairness metric would be useful.[848]

In other instances, when private sector firms use algorithms, courts should consider the financial situation, market power and size of the undertaking. It is obviously reasonable to demand public organisations, which are called to protect and promote public interests and democratic values, to forgo what may be considered to be more efficient practices from an economic perspective, in favour of increasing equality. On the contrary, said requirement will probably have to be modulated when the respondent is a private firm, especially if it does not have a large economic capacity and proves that it would have difficulties in bearing the cost of modifying the algorithm's results. It is also essential to bear in mind that there are many different types of private technological firms which design and deploy algorithms, some of which have such economic capacity and market power that have become monopolies in practice. A clear example of this is Google. It may therefore be possible to establish a lower threshold in the leniency that large technological corporations are granted when considering their interests in the proportionality analysis.

4. THE EU FRAMEWORK OF SUBSTANTIVE EQUALITY

While it is necessary to establish the ways in which cases of algorithmic discrimination can be detected and proven once they have taken place (*ex post* mechanisms), it is also important to lay out the series of measures that are specifically aimed towards preventing algorithmic discrimination built from the perspective of redressing systemic group-based inequalities.

The idea of teaching notions of equality and fairness to algorithms in order to prevent instances of algorithmic discrimination has been brought forward in a significant number of academic papers and institutional reports.[849] The growing body of literature focused on

[848] DWORK, C. *et al.*, "Fairness through awareness", *cit*, 2012, pp. 214-226.
[849] BORNSTEIN, S., "Antidiscriminatory algorithms", *cit.*, 2019, pp. 519-572; CORBETT-DAVIES, S. & GOEL, S., "The measure and mismeasure of fairness…", *cit.*, 2018, pp. 1-25; DWORK, C. *et al.*, "Fairness through awareness", *cit.*, 2012, pp. 214-226; FLANAGAN, M., HOWE, D. C., & NISSENBAUM, H., "Embodying values in technology: theory and practice", in VAN DEN HOVEN, J. & WECKETT, J., (eds.), *Information Technology and Moral Philosophy,* Cambridge, Cambridge Unviersity Press, 2008, pp. 322-353; HACKER, P., "Teaching fairness

creating "fair" algorithms approaches this topic with two objectives: 1) detecting and removing the discriminatory patterns produced by automated systems and, 2) creating non-discriminatory algorithms.[850] In addition, a growing discussion regarding different conceptualisations of fairness has also emerged.[851] Within the framework of algorithmic discrimination, the tensions that exist between different notions of equality become especially visible due to the fact that these systems require making very specific choices when they are being developed.[852]

The following pages briefly analyse different types of measures aimed towards achieving substantive equality. Special attention is paid to positive action measures for they are particularly controversial, especially when analysed from traditional liberal perspectives. Other mechanisms, such as promotion of equality and mainstreaming will also be referred to. Once said analysis is carried out, this chapter specifically focuses on the different mechanisms that have been suggested by the literature on algorithmic fairness in order to prevent and deal with instances of discrimination; a succinct assessment of the legality of said measures is also developed.

4.1. REGULATION AND POLICY FOR SUBSTANTIVE EQUALITY

4.1.1. Positive or affirmative action

4.1.1.1. Concept

Positive or affirmative action is a proactive policy that is designed and implemented by public and private organisations with the objective of compensating for the structural

to artificial intelligence…", *cit.*, 2018, pp. 1143-1186; KOCHI, E., "How to prevent discriminatory outcomes in machine learning", World Economic Forum Global Future Council on Human Rights, 2018.; KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", European Parliamentary Research Service, 2019; LOFTUS, J. *et al.*, "Causal reasoning for algorithmic fairness", *cit*., 2018; YANISKY-RAVID, S. & HALLISEY, S., "Equality and privacy by design: a new model of artificial intelligence data transparency via auditing, certification, and safe harbor regimes", *Fordham Urban Law Journal*, vol. 46, No. 2, 2019, pp. 428-486; WILLIAMS, B. A., BROOKS, C. F. & SHMARGAD, Y., "How algorithms discriminate based on data they lack: challenges, solutions, and policy implications", *Journal of Information Policy*, No. 8, 2018, pp. 78-115; ŽLIOBAITĖ, I. & CUSTERS, B., "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models", *cit.*, 2016, pp. 183-201.

[850] DWORK, C. *et al.*, "Fairness through awareness", *cit*, 2012, pp. 214-226; FELDMAN, M. *et al.*, "Certifying and removing disparate impact", cit., 2015, pp. 259-268.

[851] CORBETT-DAVIES, S. & GOEL, S., "The measure and mismeasure of fairness…", *cit.*, 2018, pp. 1-25

[852] KROLL, J. A., "Accountable algorithms", *cit.*, 2017, pp. 695-696.

discrimination suffered by certain groups.[853] The underlying logic behind affirmative action policies is that the persistence of dominant narratives and the systemic discrimination suffered by traditionally oppressed groups leads to members of said groups undergoing discrimination even when formal equal opportunities exist.[854]

4.1.1.2. General justifications

i) *The inherent goodness of a more equal society*

DWORKIN focused on university admissions that set quotas for non-white students in order to develop his argument for affirmative action policies. DWORKIN justifies affirmative action on the basis that it produces a larger benefit for the community as a whole. This benefit is, however, not framed within utilitarianism (or efficiency) but within idealism in the sense that affirmative action policies draw us closer to an ideal society in which all individuals are equally and fully integrated.[855]

ii) *Incorporating the specificities of disadvantaged groups for a fully democratic society*

While DWORKIN's argument could be employed generally to defend any affirmative action policy, this does not mean that affirmative action measures in all areas of society can or must be justified, but that the incorporation of said measures should be considered in those areas or sectors, such as education and political representation, which provide individuals with the tools for effective participation and inclusion in social power structures. In this sense, other arguments have been developed in order to defend and justify specific positive action measures.

One of the most significant, and most debated, affirmative action policies undertaken in order to include the members of disadvantaged groups in societies' power institutions have been gender quotas in legislative bodies, either through reserved seats in parliament, political party or legislated candidate quotas.[856] RODRÍGUEZ RUÍZ and RUBIO MARÍN develop an argument in favour of parity democracy built on the differentiated notions of femininity and masculinity persistent in society. They argue that in order for (female) values associated to the private

[853] CROSBY, F. J., IYER, A. & SINCHARO, S., "Understanding affirmative action", *Annual Review of Psychology*, vol. 57, 2006, p. 587.
[854] *Idem*, p. 600.
[855] DWORKIN, R., *Taking Rights Seriously*, Cambridge (Massachusetts), Harvard University Press, 1977, p. 232.
[856] IDEA, "Gender quotas database", 2019. Available on 28th April 2019 at: https://www.idea.int/

sphere to fully enter the (male-dominated) public sphere, female representation is necessary, and thus, the entry of values attributed to women that will aid in shifting the masculine values power structures are built on.[857]

All in all, positive action tools are justified on the basis of structural discrimination and are aimed towards correcting and redressing the deficiencies generated by a social structure of intergroup inequality that places members of certain groups at a general social disadvantage.[858]

4.1.1.3. Specific tools and mandates

Other than quotas, affirmative action policies in general include taking into special consideration disadvantaged group membership and compensating structural discrimination. For example, a form of affirmative action would be providing women who have had children "extra points" in employment application processes. This additional valuation would not be given to men with children. The measure would be designed to consider the reality that women interrupt their careers at a higher frequency than men due to the fact that caring responsibilities still largely fall on them. However, these types of measures must be implemented with great care given that they can lead to counterproductive results and further reinforce existing gender roles.

There are also certain types of measures that are aimed towards adapting to the specific needs of members of protected groups.[859] Obligations to adjust to the specific needs of certain individuals have especially been set forth in order to protect individuals with disabilities and promote their social inclusion. These measures are generally encompassed under the notion of "reasonable accommodation".[860]

Since the obligation of reasonable accommodation has been mainly developed from the perspective of the discrimination suffered by individuals with disabilities, said concept is only included in the Employment Equality Directive and exclusively in reference to individuals with disabilities. Reasonable accommodation requires making adjustments in

[857] RODRÍGUEZ RUÍZ, B. & RUBIO MARÍN, R., "Constitutional justification of parity democracy", *cit*., 2009, pp.1180-1183.
[858] BARRÈRE UNZUETA, M. A., "Igualdad y 'discriminación positiva'…", *cit.*, 2003b, pp. 18-20.
[859] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", *cit.*, 2018, pp. 69-70.
[860] FREDMAN, S., *Discrimination Law*, *cit.*, 2011, p. 154.

order to ensure the full participation and integration of individuals with disabilities. For example, in the workplace context it "means that employers shall take appropriate measures, where needed in a particular case, to enable a person with a disability to have access to, participate in, or advance in employment, or to undergo training" as long as said measures do not impose a "disproportionate burden" on the employer.[861]

Reasonable accommodation is therefore presented as a remedy against indirect discrimination. As it was already indicated, indirect discrimination takes place when apparently neutral provisions result in a disadvantage for members of a sub-category within a protected ground, generally, members of historically subordinated groups. Hence, by establishing reasonable accommodation mandates, the EU anti-discrimination framework recognises the existence of specificities that characterise members of certain groups; the fact that those specificities are not taken into consideration when certain rules and social norms are developed and the need for adapting to those specificities.

This mandate is only explicitly mentioned with regard to the protection of individuals with disabilities due to the fact that the disadvantage of these individuals is far more obvious and, in many ways, pervasive, than the disadvantage that structural discrimination generates for members of other protected groups. However, the fact that the duty of reasonable accommodation is only considered with regard to individuals with disabilities, shows how the EU anti-discrimination framework does not fully incorporate or acknowledge the fact that social norms and institutions are designed from the perspective of historically dominant groups, thereby hampering the chances of full social participation of members of all disadvantaged groups. Nonetheless, the notion of "reasonable accommodation" is also sometimes employed in the context of religious discrimination.[862] Therefore, although not legally mandated, this duty can also be introduced with regard to other protected groups.

4.1.1.4. Conflicts

Positive action measures can enter into conflict with a number of principles and rights. In particular, when they specifically prioritise members of the protected group over members of the advantaged group, a trade-off between formal and material equality necessarily takes

[861] Article 5 of the Employment Equality Directive.
[862] VICKERS, L., "Religion and belief discrimination in employment – the EU Law", Luxembourg, Office of Publications of the European Communities, 2006, pp. 19-23.

place. It is therefore vital for justifications to these types of actions to be structured through a carefully crafted proportionality analysis.

Additionally, affirmative action policies sometimes can be counterproductive and lead to undesired results, such as assimilation of the disadvantaged group to the advantaged group without considering the particularities of the former. For instance, political parties should take into consideration that most caring responsibilities within households fall on women. Otherwise, even if there is an equal proportion of men and women in representative positions, a wide array of situations typical of political organisations end up excluding women from real power and decision-making. For instance, when meetings are held at times in which women generally exercise their caring activities, such as late in the afternoon, or in informal settings, women do not attend and are therefore left out of highly important and strategic decisions.

These measures also risk reinforcing inferiority stereotypes by conveying the idea that the members of disadvantaged groups are being given something they do not naturally deserve. It is therefore essential for the results of affirmative action policies to be tested and for these measures to be set up in combination with policies for the promotion of equality in a more general sense. In this sense, affirmative action measures have proven to be effective, and even necessary, in several contexts, including in increasing female political representation and engagement as well as their presence in company boardrooms.[863]

The conflicts that arise regarding the rights of dominant groups (i.e.: white men)[864] when discussing affirmative or positive action measures stem from setting the notions of formal equality and non-discrimination as the main categories of analysis. The problem then is: if we do nothing, disadvantaged groups are discriminated against because social structures are designed from the perspective of dominant groups and narratives; however, if we implement measures that privilege members of disadvantaged groups, we are discriminating against non-

[863] Resolution 2111 (2016) on assessing the impact of measures to improve women's political representation of the Council of Europe's Parliamentary Assembly; MORGENROTH, T. & RYAN, M. K., "Quotas and affirmative action: understanding group-based outcomes and attitudes", *Social and Personality Psychology Compass*, vol. 12, No. 3, 2018, pp. 1-14; NORDEN, "Gender equality – The nordic way", Copenhaguen, Nordic Council of Ministers, 2010.

[864] Dominant groups do not just include being white and being male but a long list of characteristics that have been examined throughout the dissertation but that it is not necessary to repeat here.

members, hence explicitly breaching equality and non-discrimination as core principles of justice.[865]

As it has already been indicated throughout the dissertation, strictly speaking, discrimination can mean the differential treatment of both members of disadvantaged and members of dominant groups. Although this dissertation approaches discrimination as the result of the historical and existing subordination suffered by some social groups, it is easy for legal and political analysts to fall back on the previously mentioned more straightforward, narrow and formal concept of discrimination. In order to properly understand why affirmative action can be and is justified, it is necessary to shift focus from (formal) equality and discrimination to oppression and subordination.[866]

4.1.1.5. Similarities and differences between indirect discrimination and affirmative action

Establishing the differences between indirect discrimination and affirmative action can sometimes present significant difficulties. After all, they are both built upon similar approaches to inequality. Both affirmative action and indirect discrimination address criteria or provisions that formally respect the right to equality for they are apparently neutral, they acknowledge and focus on the group-dimension of discrimination and both are mechanisms of reaction against the differentiated social standing of identity or population groups.[867] Additionally, determining that an apparently neutral practice or criterion is in fact discriminatory entails the need to shift said practice or criterion in order to accommodate the particularities of the disadvantaged group.

The first difference is that, while indirect discrimination is a category of analysis used in the judgment of discrimination cases, affirmative action is a type of policy, although it must obviously respect the legal framework within which it is developed and implemented. In addition, indirect discrimination does not, in principle, aim to provide a less favourable treatment to members of the advantaged group but to ensure that apparently neutral practices do not only formally respect the right to equality, but also do so from a material perspective. Conversely, affirmative action is justified on a series of historical and present injustices that

[865] YOUNG, I. M., *Justice and the Politics of Difference*, *cit.*, 1990, pp. 194-195.
[866] *Idem*, pp. 195-196.
[867] AGUILERA RULL, A., *Contratación y Diferencia…*, *cit.*, 2013, p. 183.

have led to group disadvantage and aims to compensate for said injustices, in some cases, by establishing measures that prioritise and privilege members of the disadvantaged group.[868]

Nonetheless, there are cases in which the lines between indirect discrimination and affirmative action become especially blurry. For example, if a court determines that the group parity metric at a 40 to 60% representation ratio for each sex as the non-discrimination and equality rule for the preselection in a recruitment process. If the algorithm, for apparently objective reasons had preselected more than 60% of male candidates it would be deemed discriminatory. Hence, algorithmic affirmative action would have to be introduced in the system in order to ensure that the preselection process complies with the equality and non-discrimination mandate.

4.1.1.6. EU framework and case law

i) *Legal framework*

All of the EU Equality Directives contain provisions that allow member states to adopt measures of positive action.[869] The TFEU also contains a specific provision establishing that member states can adopt "measures providing for specific advantages in order to make it easier for the underrepresented sex to pursue a vocational activity or to prevent or compensate for disadvantages in professional careers".[870] In addition, recommendation on the Promotion of Positive Action for Women was adopted by the Council of the EU in December 1984.

ii) *CJEU case law*

While the CJEU has not ruled on positive action measures in the context of equality in access to goods and services, there are several rulings that cover different affirmative action measures in the area of employment. The types of positive actions that have generated controversies in employment cover measures to enhance female work-life balance and mechanisms to favour access to employment of members of a disadvantaged group, mainly women.

[868] *Idem*, p. 184.
[869] Art. 5 Racial Equality Directive; art. 3 Gender Employment Equality Directive; art. 6 Gender Goods and Services Directive and art. 7 of the Employment Equality Directive.
[870] Article 157.4 TFEU.

The Court has not been lenient in accepting positive action measures that prioritise "the under-represented sex" (generally women) in access to employment, promotion and training. In the *Kalanke* case, the Court ruled that if two individuals that are shortlisted to occupy a job are equally qualified, automatic priority can not be given to the under-represented sex.[871] Conversely, the Court did deem that a similar measure was not contrary to the provisions to the Gender Equality in Employment Directive in the *Marschall* case due to the fact that in this instance the positive action measure contained a "saving clause" according to which, when a male and female candidate were equally qualified, the female candidate should be prioritised "unless reasons specific to an individual [male] candidate tilt the balance in his favour".[872] It was after the *Marschall* ruling that the provision included in the TFEU (the then Treaty establishing the European Community [TEC]) allowing member states to adopt positive action measures to compensate for the disadvantages suffered by women in their professional careers was adopted.

Similarly to the *Marschall* ruling, the Court established in the *Badeck* case that a series of actions designed to provide specific advantage to women in access, training and promotion in public employment in Germany were lawful as the measures contained sufficient flexibility in order to consider the specific situations of all candidates and did not grant women an automatic advantage.[873]

Finally, in the *Abrahamsson* case, the Court had to decide on a provision that automatically prioritised women over men as long as they complied with the necessary qualifications for a position even if the male candidate was more qualified. The CJEU deemed said provision to be contrary to the TEC and the Gender Equality in Employment Directive for it was "disproportionate to the aim pursued".[874]

In this vein, the CJEU seems to generally differentiate between measures that it considers to be framed within the idea of positive discrimination and measures that can be classified as

[871] CJEU Judgment 17th October 1995, C-450/93, Eckhard Kalanke v. Freie Hansestadt Bremen, paragraphs 22 and 23: "National rules which guarantee women absolute and unconditional priority for appointment or promotion go beyond promoting equal opportunities and overstep the limits of the exception in Article 2(4) of the Directive. Furthermore, in so far as it seeks to achieve equal representation of men and women in all grades and levels within a department, such a system substitutes for equality of opportunity as envisaged in Article 2(4) the result which is only to be arrived at by providing such equality of opportunity".
[872] CJEU Judgment, 11th November 1997, C-409/95, Hellmut Marschall v. Land Nordrhein Westfalen, operative part.
[873] CJEU Judgment 28th March 2000, C-158/97, Badeck v. Landesanwalt beim Staatsgerichtshof des Landes Hessen.
[874] CJEU Judgment 6th July 2000, C-407-98, Abrahamsson v. Fogelqvist, paragraph 55.

positive or affirmative action. The former are those that set automatic priorities for the disadvantaged group and are therefore contrary to the Court's interpretation of the right to equality and non-discrimination. The latter are designed more as measures of outreach but that do not automatically privilege the disadvantaged group and are therefore tolerable from the perspective of formal equality.

In the rulings on positive actions regarding work-life balance, the Court has generally recognised the fact that measures should be taken in order to ensure that the caring responsibilities that mostly fall on women do not trump their professional careers. However, it has addressed each specific case taking into consideration the fact that providing very significant work-life advantages to women may in fact be counterproductive as it can help perpetuate traditional gender roles.[875]

iii)*ECHR case law*

The ECHR bases its rulings on positive action measures on article 14 and Protocol 12 of the European Convention on Human Rights, which set out the prohibition of discrimination. The Court has considered that the referred rules allow states to adopt policies that entail an unequal treatment between members and non-members of protected groups or that are specifically accommodated to the particularities of disadvantaged groups, in order to correct for existing inequalities.

For instance, *Andrle v. Czech Republic*, the ECHR ruled that lowering the pensionable age of women who had raised children did respect article 14 of the Convention as this measure took into consideration the lower salaries and pensions that women receive, on average, and the fact that the existing social reality in the Czech Republic meant that women were expected to care for their children and household while also working full-time, with the resulting hardship and inequality with regard to men that this entails.[876]

Moreover, if positive action measures are not adopted, since states have positive obligations to protect against discrimination, they can be deemed to be in breach of the prohibition to

[875] CJEU Judgments 19th March 2002, C-476/99, H. Lommers v. Minister van Landbouw, Natuurbeheer en Visserij; 30th 2010, C-104/09, Pedro Manuel Roca Álvarez v. Sesa Start España ETT SA; and, 16th July 2015, C-222/14, Konstantinos Maïstrellis v. Ypourgos Dikaiosynis, Diafaneias kai Anthropinon Dikaiomaton.
[876] ECHR Judgment 17th February 2011, 6268/08, Andrle v. Czech Republic, paragraphs 53 and 55.

discriminate.[877] In addition, there are several cases in which article 14 of the Convention has been triggered as a result of the lack of action by states in taking into consideration the particularities of the situations endured by members of protected groups. For instance, in *Abdu v. Bulgaria*, the Court established the need to judge violent crimes motivated by racist attitudes in a different manner than other violent crimes, that is, considering the special social harm caused by the underlying racist motive. Not doing so, the Court determined, was constitutive of an infringement of article 14 of the Convention.[878]

4.1.2. Mainstreaming and promotion of equality

Existing approaches to discrimination generally fail to provide solutions that are built based on a comprehensive analysis of the structural discrimination that is present in society, leading to the creation of regulatory instruments and the application of policies that only partially tackle historical discriminatory structures.

Approaching situations of discrimination from the perspective of the disadvantage caused to members of historically subordinated groups is necessary both when applying non-discrimination law as well as when designing general policies and specific measures aimed towards achieving equality and fighting discrimination. The specific tool that is used in order to ensure that policies are not biased, thus providing privileges to dominant sectors, is non-discrimination mainstreaming. Said tool should be used and implemented through intersectional perspectives.

---

[877] ECHR Judgment 30th June 2016, 51362/09, Taddeucci and Mccall v. Italy, paragraph 81: "According to the Court's well-established case-law, in order for an issue to arise under Article 14 there must be a difference in treatment of persons in relevantly similar situations (see Hämäläinen, cited above, § 108), or an issue will arise when states fail to treat differently persons whose situations are significantly different (see Thlimmenos, cited above, § 44 in fine). On the latter point the Court reiterates that Article 14 does not prohibit a member state from treating groups differently in order to correct "factual inequalities" between them; indeed in certain circumstances a failure to attempt to correct inequality through different treatment may in itself give rise to a breach of the Article (see Case "relating to certain aspects of the laws on the use of languages in education in Belgium" (merits), 23 July 1968, § 10, Series A no. 6; Stec and Others v. the United Kingdom [GC], nos. 65731/01 and 65900/01, § 51, ECHR 2006-VI; and Muñoz Diaz v. Spain, no. 49151/07, § 48, ECHR 2009). Furthermore, the Court has already accepted in previous cases that where a general policy or measure has disproportionately prejudicial effects on a particular group, it is not excluded that this may be considered as discriminatory notwithstanding that it is not specifically aimed or directed at that group and there is no discriminatory intent. Such a situation may amount to "indirect discrimination". This is only the case, however, if such policy or measure has no "objective and reasonable" justification (see, among other authorities, Baio v. Denmark [GC], no. 38590/10, § 91, 26 May 2016; S.A.S. v. France [GC], no. 43835/11, § 161, ECHR 2014 (extracts); D.H. and Others v. the Czech Republic [GC], no. 57325/00, § 184, ECHR 2007-IV; and Hugh Jordan v. the United Kingdom, no. 24746/94, § 154, 4 May 2001)."

[878] ECHR Judgment 11th March 2014, 26827/08, Abdu v. Bulgaria.

Non-discrimination mainstreaming aims to include the particularities and specific needs of traditionally oppressed groups in all decision-making through addressing all issues from the perspective of equality.[879] For example, a typical way in which structural discrimination is reinforced is by placing the experiences of men at the centre of decision-making processes, therefore resulting in apparently neutral decisions that do in fact harm women. For example, returning to the political party meeting example that was mentioned earlier on, if a meeting in a public or private organisation is set late in the afternoon, women, who still take up most responsibilities within the household, will find more difficulties to attend. The objective of non-discrimination mainstreaming is to introduce the experiences and values of all disadvantaged groups thereby leading to decisions that can accommodate traditionally oppressed groups and that do not place white males of a certain socioeconomic status as the stereotypical individual upon whom rights are constructed.

Non-discrimination mainstreaming, or rather, gender mainstreaming in this case, is for instance essential in order to reduce the gender pay gap. The Gender Employment Equality Directive states in article for that "where a job classification system is used for determining pay, it shall be based on the same criteria for both men and women and so drawn up as to exclude any discrimination on grounds of sex". As the previously cited Spanish Constitutional Court Judgment No. 145/1991 showed, differences are in many cases the result of providing traditional male attributes, such as physical strength, with a higher value than traditional female attributes. By taking into consideration the particularities of women and how the jobs that they occupy are aligned in many cases with traits associated to women, the equal evaluation of jobs that are mainly developed by men and jobs that are mainly developed by women would be advanced.

It is important to highlight that only the Gender Employment Equality Directive dedicates a whole article to mainstreaming, as article 29 establishes that "member states shall actively take into account the objective of equality between men and women when formulating and implementing laws, regulations, administrative provisions, policies and activities in the areas referred to in this Directive". Furthermore, the Race and Employment Directives only

[879] DG JUSTICE "Compendium of practice on non-discrimination/equality mainstreaming", European Commission, 2011. Available on 28th April 2019 at: https://publications.europa.eu/

mention mainstreaming in the context of assessing the impact of measures on women and men.[880] The Gender Goods and Services Directive does not mention mainstreaming.

Finally, it is also important to consider measures aimed towards the promotion of equality that are not comprehended under the scope of affirmative action and therefore focus on achieving equality of opportunity rather than equality of results. An example of this type of measure is creating specific spaces in workplaces that only women who are breast-feeding can use. These measures sometimes overlap with "reasonable accommodation" mandates but fall outside of what can be considered affirmative action in the strict sense.

4.2. PROPOSALS AND POSSIBILITIES FOR ALGORITHMIC SUBSTANTIVE EQUALITY

4.2.1. Substantive equality in the tech sector

Lack of diversity in the tech sector has been pointed out as one of the issues that lies at the origin of algorithmic discrimination. One of the possible policy choices that could be made in order to advance algorithmic equality is enforcing diversity in the tech sector. However, since said type of intervention requires setting important limits to the freedom to conduct a business it will probably face a general backlash and deemed contrary to the limits of positive action measures under current interpretations of the EU equality and anti-discrimination framework.

Nonetheless, considering the way in which the CJEU has interpreted positive action measures in the field of employment it may be possible to introduce measures that prioritise hiring, training and promoting women and racial minorities as long as an automatic advantage is not granted to them.

4.2.2. Substantive algorithmic equality. Aspirations of equality and their legal standing

4.2.2.1. Equality by design

The GDPR contains a provision that mandates all data processing systems to include "data protection by design and by default",[881] which requires controllers to establish measures and

[880] Article 17.2 Race Equality Directive and article 19.2 of the Employment Equality Directive.
[881] Article 25 GPDR: "1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and

safeguards that ensure full compliance with the Regulation. These measures have to be applied "both at the time of the determination of the means for processing and at the time of the processing itself". In this vein, several scholars have suggested implementing "equality by design"[882] as a principle and mandate that should apply in the development of algorithms and which entails "the structuring of [algorithms] in a manner that is sensitive to prevailing forms of … discrimination" [883] and which ensures that they are designed in ways that take into consideration the existence of social structures of discrimination and the ways in which these structures systematically situate the members of certain population groups in a disadvantaged position. Equality by design therefore aims to detect instances of discrimination, correct them and, in some cases, incorporate notions of systemic discrimination to correct social imbalances between members of advantaged and disadvantaged groups.

The tensions between different notions of fairness and equality have become especially present when it comes to algorithmic equality, and the discussion concerning them brings about questions regarding the extent to which regulatory instruments should articulate more precise definitions of the rights to equality and non-discrimination when it comes to algorithmic design. This discussion is especially relevant from a legal perspective, as it materialises the trade-off between equality and efficiency, and forces regulators to establish very specific objectives regarding whether they want to frame automated processes from the perspective of formal or substantive equality and, in the second case, whether equality of opportunity or of results should be prioritised. It is also relevant to point out that claims for more specific provisions that stipulate how far traditional forms of affirmative action can go, have also been expressed.[884]

---

organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects. // 2. The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons."

[882] HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, p. 1146; RENAN BARZILAY, A. & BEN-DAVID, A., "Platform inequality…", *cit.*, 2017, pp. 427-431; XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *cit*., 2020, p. 179. YANISKY-RAVID, S. & HALLISEY, S., "Equality and privacy by design…", *cit.*, 2019, pp. 428-486.

[883] RENAN BARZILAY, A. & BEN-DAVID, A., "Platform inequality…", *cit.*, 2017, p. 430.

[884] ELLIS, E. & WATSON, P., *EU Anti-discrimination Law*, Oxford, Oxford University Press, 2nd ed., pp. 505-506.

When the source of inequality lies within structural discrimination and the discriminatory results are shaped in the form of accurate statistical discrimination,[885] the tension between different conceptualisations of fairness becomes especially relevant. As it was already explained in chapter II, statistical discrimination results from prejudices held by individuals that can be both accurate and inaccurate. Inaccurate beliefs and biases held regarding certain traditionally oppressed group stem from mental shortcuts (heuristics) taken by individuals when analysing people or situations or of a lack of information regarding certain population groups.[886] This form of discrimination does not present many problems on a theoretical level as long as it is inaccurate. Conversely, instances of accurate statistical discrimination, that is, cases in which decisions based on prejudices regarding members of a traditionally oppressed group are accurate, generate significant theoretical and practical tensions. When statistical discrimination actually reflects the behaviour or situation of a certain group, questions must be raised on whether using these methods is legitimate or whether notions of equality should be prioritised.

For instance, coming back to the scoring example that was discussed in chapter III. It is important to keep in mind that individual scores are no longer used exclusively in finance but also in many other areas, however, for the purposes of the point made the following example will specifically focus on the use of algorithms to determine credit scores. It may seem completely logical for the algorithm to draw on an individual's employment data in order to determine their creditworthiness. Consequently, employees with lower salaries will be assigned lower scores. Considering the fact that the gender pay gap is an undeniable reality,[887] this supposedly rational discrimination does nothing but help the persistence of economic gender discrimination.

In cases of accurate algorithmic discrimination, making choices regarding what perspective of equality and non-discrimination should be adopted are especially difficult since producing non-discriminatory algorithms will entail introducing affirmative action.

The following pages set out the different (and sometimes conflicting) definitions of fairness and equality that are being used both by the literature on algorithmic fairness and in practice

[885] BOHREN, J. A. *et al.*, "Inaccurate statistical discrimination", *cit.*, 2019.
[886] *Idem*, p. 3.
[887] EU COMMISSION, "The gender pay gap situation in the EU", 2018. Available on 22nd February 2019 at: https://ec.europa.eu/; IMF, "Pursuing women's economic empowerment", May 31st 2018. Available on February 22nd 2018 at: https://www.imf.org/

by system developers. The approaches discussed below are either built from the perspective of formal or substantive equality. There is not one single valid solution to prevent and deal with algorithmic discrimination and all perspectives can be useful to a certain degree. However, the structural dimension of discrimination must be kept in mind at all times when deciding how "equality by design" will be implemented in each case, which means that all versions of equality and fairness will not be equally applicable to all cases and that the approaches that prioritise substantive equality should always be at least considered.

4.2.2.2. Individual and group fairness

Individual and group fairness were already addressed in the sections regarding direct and indirect algorithmic discrimination. These notions of equality (or fairness) can be particularly useful in detecting instances of discrimination. However, they can also serve as the framework within which to design algorithmic positive action. That is, depending on the definition of fairness and aspirations of equality we deem desirable, the level of affirmative action we consider justifiable will shift.

i) *Individual fairness*

Individual fairness focuses on the distance between individual subjects and do not take into consideration the common elements, other than protected group membership, that can appear in members of disadvantaged groups. For example, it compares the outcomes of two individuals and their characteristics at a given point in time and space and, as long as there are elements that differentiate them that are not protected group membership and which are relevant to the decision being made, the decision is considered fair.

Individual fairness perspectives therefore fail to consider the social elements that may have influenced the particular characteristics of each individual. These perspectives or metrics are therefore not useful in order to establish affirmative measures that, for instance, automatically prioritise members of a disadvantaged group in access to certain job positions even when they present worse qualifications or work experience than candidates that do not belong to a disadvantaged group.

However, individual fairness does work as a general framework upon which to build affirmative action measures that are consistent with CJEU case law on positive action. The

individual fairness measure can be used in order to determine that a man and a woman share the exact same qualities that are relevant for a job and to prioritise women as long as the "saving clause" that does not grant automatic advantage to women is put in place.

ii) *Group fairness*

Group fairness takes into consideration the existence of structural discrimination and is based on notions of substantive equality and, particularly, in measures that pursue equality of outcomes between members and non-members of protected groups.[888] As it was already indicated in the section on indirect algorithmic discrimination, group fairness can be measured and expressed as demographic parity or accuracy parity. This form of group fairness can be especially useful in contexts that help to enhance full access and participation of members of disadvantaged groups in social power structures, such as political representation and higher education. For instance, group fairness as a nuanced form of demographic parity is the framework used in quota systems that establish a minimum and maximum percentage of representatives of each sex. Accuracy parity is useful in those contexts in which full demographic parity or even the nuanced form of demographic parity cannot be acquired, such as credit or recidivism scoring systems.

Demographic parity is currently generally accepted and materialised in several countries in Europe in political representation and, in some cases, even as a mandate to determine the composition of private organisations' executive boards. Statistical parity may also be useful in employment and higher education pre-selection processes. We consider accuracy parity to also provide a useful measure to determine the acceptability of certain affirmative action measures, especially when they apparently clash with notions of efficiency.

However, statistical or demographic parity can generate very unfair results at the individual level and result in outcomes that are contrary to public interests, especially in its most restrictive versions. For example, taken to an extreme, statistical parity could justify incarcerating equally representative amounts of Hispanic, black and white individuals even if that meant sending innocent individuals to prison.[889]

[888] FRIEDLER, S. A. *et al.*, "On the (im)possibility of fairness", *cit.*, 2016, pp. 11-14.
[889] BERK, R. *et al.*, "Fairness in criminal justice risk assessments…", *cit.*, 2018, p. 14.

4.2.2.3. The combination of individual and group fairness

DWORK *et al.* try to solve the conflict between equality and efficiency within algorithms by providing a theoretical example of how an organisation using targeted advertising can achieve efficient results by carrying out classifications of possible future customers in which protected group membership is not relevant.[890] The objective of this paper is to prevent the use of predatory advertising, which is used by different types of companies, such as private colleges or credit card firms, to sell low quality or toxic products and services to individuals who are in vulnerable situations and who are generally members of disadvantaged groups.[891]

DWORK *et al.* try to achieve this objective by developing a measure of similarity between the individuals whose data is analysed and used in order to develop systems of targeted advertising.[892] They base their work on the notion of individual fairness[893] and therefore contend that individuals who have a similar measure of, for example, creditworthiness, should be treated similarly.[894] This notion of fairness is basic and generally not contested from the perspective of traditional conceptualisations of freedom and efficiency as it does ensure procedural regularity and does not delve into the issues that arise from historical group oppression. The paper does however suggest the possibility that in some instances measures of group fairness should also be introduced in order to correct for ongoing social biases.[895]

Their work mainly focuses on individual fairness due to the fact that applying notions of group fairness to targeted advertising (and many other uses of algorithms) is not appropriate. For example, it is possible that when addressing a fraction of a traditionally oppressed group for a certain purpose, the wrong subset is chosen due to the fact that the parameters used are those designed according to the values of the dominant group, and those differ to the values of the traditionally oppressed group.[896] It is also possible that the classification or selection algorithm is purposefully designed so that the subset of the traditionally discriminated group

[890] DWORK, C. *et al.*, "Fairness through awareness", *cit*, 2012, p. 214.
[891] US SENATE COMMITTEE ON COMMERCE, SCIENCE & TRANSPORTATION, MAJORITY STAFF, "A review of the data broker industry…", *cit.*, 2013, p. i.
[892] DWORK, C. *et al.*, "Fairness through awareness", *cit*, 2012, p. 214.
[893] KROLL, J. A. *et al.*, "Accountable algorithms", *cit*., 2017, p. 685.
[894] DWORK, C. *et al.*, "Fairness through awareness", *cit*, 2012, p. 215.
[895] *Ibidem.*
[896] DWORK, C. *et al.*, "Fairness through awareness", *cit*, 2012, p. 218: "Suppose in the culture of S the most talented students are steered toward science and engineering and the less talented are steered toward finance, while in the culture of S^c the situation is reversed: the most talented are steered toward finance and those with less talent are steered toward engineering. An organization ignorant of the culture of S and seeking the most talented people may select for "economics", arguably choosing the wrong subset of S, even when maintaining parity".

that is selected does not fit the criteria and that, for example, the candidates of a racial minority group that are selected for interviews are those that do not have the degree of education required for a position so that not hiring people form that minority group would be justified and the firm cannot be accused of not carrying out an initial selection with statistical parity.[897] Similarly, predatory ads can be targeted to the same percentage of each group but ensuring that the members of the dominant group that are targeted will not be interested in the product or service.[898]

Given the limitations of conceptualising fairness as statistical parity, DWORK *et al.* mainly base their proposal on individual fairness but constrained by the notion of "fair affirmative action".[899] This is carried out by limiting the number of members of disadvantaged groups that are assigned classifications that will result in negative outcomes for them.[900] For instance, in the case of credit scoring, the initial premise that would be considered if DWORK *et al.*'s system were used would be that within each classifier (low, medium and high credit scores) its members would be treated in a similar manner. Additionally, in order to include fair affirmative action, the algorithm would limit the amount of members of racial minorities classified with a low credit score. DWORK *et al.* prove that doing so offers results that are almost as accurate and good as those offered by algorithms to which fairness has not been incorporated.[901]

4.2.2.4. Combining individual fairness with randomisation

KROLL *et al.*[902] propose introducing an element of randomness into the algorithm with the objective of validating the system. Hence, they argue, if an algorithm is inadvertently using a protected characteristic, such as sex, to predict job performance, and it therefore mainly hires men, the algorithm once it tests actual performance against its predictions, will probably end up associating good performance with variables that are proxies for being male. They therefore suggest that the algorithm should randomly recommend that some candidates that are not predicted to do well get hired so that the system will be able to reduce the possibility that on-going testing simply works by reinforcing the assumption that certain traits associated to men accurately predict good job performance. This will also help to ensure that the

[897] DWORK, C. *et al.*, "Fairness through awareness", *cit*, 2012, p. 218.
[898] *Ibidem*.
[899] *Idem*, pp. 215, 220.
[900] KROLL, J. A. *et al.*, "Accountable algorithms", *cit*., 2017, p. 687.
[901] *Ibidem*.
[902] KROLL, J. A., "Accountable algorithms", *cit.*, 2017, pp. 683-684.

algorithm is more faithful to the real world in which it is deployed than to the training and test data sets that helped develop it. By ensuring that the algorithm will randomly select individuals that would, in theory, not fare well, more women could be selected and, consequently, the algorithm could learn over time that its initial assumptions were erroneous.

4.2.2.5. Algorithmic equality of opportunity: reframing values, labels and features

One of the problems with algorithmic discrimination is that, in many instances, it will not be possible to clearly determine whether it is accurate or inaccurate. The amount of opportunities that exist during the development and training phases and even after deployment for the algorithm to incorporate biases that lead to discriminatory outcomes makes it very difficult to test for actual accuracy and especially, to detect exactly at which point biases entered the system. This is why incorporating some form of group fairness and other strategies, such as a certain degree of randomisation, are so important.

However, strategies to reframe the social structures that are incorporated to algorithms must also be implemented. Systems must be designed in ways that ensure that variables that privilege dominant groups are not introduced unless they are the only accurate measure of the target variable. The types of labels chosen, the features used to define said labels and the way in which said features are measured can lead to discriminatory systems. Hence, in addition to the test and training datasets, the architectures of automated systems should be carefully reviewed. System designers should always work with two main premises: a) if an algorithm discriminates it is generally the result of framing problems from dominant perspectives and b) even if an algorithm is apparently accurate, as long as it discriminates, it is reproducing the structures of discrimination that are deeply entrenched in our societies.

4.2.3. Implementing algorithmic affirmative action and promotion of equality

The current legal framework and interpretation of positive action by the CJEU is unquestionably limited both in scope and tolerance of said type of measures. As for the ECHR, it has been generally more lenient in accepting affirmative action measures. However, considering EU member states are much more constrained by the mandates of EU law and the CJEU, the higher threshold for considering positive action measures justified set by this Court is probably more reliable in determining what kind of algorithmic affirmative action will be allowed under EU law.

As it will be further argued in the following chapter, CJEU Judgments on discrimination notably lack an appreciation of the structures of discrimination upon which societies are built. While the CJEU has in some cases analysed discrimination cases from the perspective of group disadvantage, it generally relies more heavily on notions of formal equality rather than substantive equality. From this perspective, the introduction of group parity measures in order to render algorithms non-discriminatory will be deemed to be in violation of the equality and non-discrimination framework, particularly when algorithms are proven to be accurately discriminating against protected groups. However, the mixed approaches to algorithmic fairness that do, to a certain extent, incorporate the collective and structural dimensions of discrimination, may be, if sufficiently justified as pursuing a legitimate aim and being proportionate, deemed to fall within the admissible positive action measures foreseen in the EU Equality Directives.

It is also necessary to highlight once more that, especially when comes to accurate algorithmic discrimination, determining whether an algorithm is discriminatory or not and considering that algorithmic positive action is justified fall along the same continuum. If courts consider an accurate algorithm to be discriminatory and force the respondent to change the system's parameters, this will constitute an affirmative action measure.

Finally, with regard to non-controversial measures, regulators and academia should generally focus on analysing the ways in which commonly accepted worldviews lead to the discrimination of disadvantaged groups in order to exactly determine how algorithmic equality mainstreaming and promotion of equality should be articulated and implemented. In this sense, it is necessary acknowledged that statistics and the way in which social phenomena are measured is not objective and that, what we currently generally consider to be the most accurate way in which to measure society, is simply the way in which dominant narratives have framed reality.

4.2.4. Using algorithms to detect discrimination

Finally, algorithmic systems should also be deployed in order to detect instances of discrimination and inequality, not only in other automated systems but also in general decision-making and other social processes. Algorithms may be particularly useful, for instance, in addressing instances of intersectional discrimination. Cases of intersectional discrimination are not easy to prove due to the difficulties of determining the extent to which

a particular measure affects an individual because of her multiple disadvantage as member of two or more vulnerable groups. Automated systems could help to detect and measure how belonging to two or more disadvantaged groups determines being treated or impacted in a specifically harmful way.

It is essential that we take advantage of the possibilities offered by these new technological advances in order to tackle aspects of inequality that have thus far only been patched up. Using software systems may help us to identify the way in which social structures are biased against the members of certain groups and redress said bias.

CHAPTER V. THE EQUALITY AND ANTI-DISCRIMINATION FRAMEWORK: LIMITS AND SHORTCOMINGS

The European equality and anti-discrimination framework covers a significant number of situations and grounds and contains measures that are not only aimed towards detecting and remedying instances of discrimination but also towards promoting equality. However, the rules that make up this framework and the way in which they are applied by the CJEU still fall short in fully and comprehensively addressing discriminatory practices and structures. In addition, when it comes to analysing the perpetuation of structures of discrimination through algorithms, the existing legal framework proves limited as it lacks the necessary tools to deal with decisions that, although not strictly discriminatory, can be highly harmful by procuring the persistence of group disadvantage.

This chapter focuses on said shortcomings and limitations, many of which are applicable to algorithmic and traditional discrimination cases. Said shortcomings are classified into three types: those that result from the limited scope of application of equality and non-discrimination instruments; those that result from the individualistic-formalistic construction of the rights to equality and non-discrimination and those that result from a lack of effective system of enforcement.

1. SCOPE OF APPLICATION

1.1. SUBJECT MATTER

The EU anti-discrimination Directives only focus on certain areas. While general prohibitions to discriminate appear in the European Convention on Human Rights (art. 14 and art. 1 in Protocol No. 12 to the Convention) and the Charter of Fundamental Rights of the EU, the areas that fall outside the Directives are not covered as intensely and comprehensively. Especially when certain areas are excluded from the application of a Directive, as it is the case with advertising in the Gender Goods and Services Directive,[903] it will be much harder to defend the application of prohibitions to discriminate to certain uses of algorithms.

[903] Article 3.3.

1.1.1. Self-employment

The Equality Directives that cover employment, including the Self-employment Gender Equality Directive, only focus on certain aspects of self-employment. Amongst other problematic issues this means that the equal pay principle does not apply to genuinely self-employed workers.[904] This could be especially problematic with regard to platform workers seeing as female platform workers generally get paid less than men whether as a result of consumer's biases, structural discrimination or a combination of both.[905]

1.1.2. Advertising

Another area that generates doubts is advertising. The Gender Goods and Services Directive explicitly excludes media and advertising from its scope of application. The Race Equality Directive does not explicitly mention advertising meaning cases of discrimination in advertising can be included under the equality protection it offers.[906]

There are three general ways in which algorithmic discrimination in advertising can take place. Individuals in which protected grounds concur, or who are somehow related to protected groups, can be discriminated by not being shown advertisements for goods and services. This can happen as a result of direct and explicit discrimination, as it was the case with Facebook's housing advertisements and the possibility that the social network offered its users from excluding certain ethnic groups.[907] This particular form of algorithmic discrimination may also fit within the definition of indirect discrimination when the algorithm discriminates on the basis of a variable or set of variables that is correlated with a protected ground.

If this particular form of advertising takes place, it could be argued that, by excluding protected groups from viewing certain goods and services, their access to them is being curtailed, hence falling within the scope of the Gender Goods and Services Directive. However, most of the literature studying this particular issue has concluded that, by excluding advertising from the scope of application of the Gender Goods and Services

[904] KÜLLMANN, M., "Platform work, algorithmic decision-making, and eu gender equality law", *cit.*, 2018, pp. 16-17.
[905] RENAN BARZILAY, A. & BEN-DAVID, A., "Platform inequality…", *cit.*, 2017; COOK, C. *et al.*, "The gender earnings gap in the gig economy…", *cit.*, 2019, pp. 1-62.
[906] XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *cit.*, 2020, p. 167.
[907] ANGWIN, J. & PARRIS JR., T., "Facebook lets advertisers exclude users by race", *cit.*, 2016.

Directive, the European legislator allows advertisers to segregate ads and use gender to target advertisements.[908]

Another way in which algorithmic discrimination can take place through advertising is through predatory advertising, that is, by specifically targeting low quality goods and services to members of protected groups and, in particular, socioeconomically vulnerable populations.[909] Even if the notion that targeting advertisements to certain population groups means limiting access to goods and services was accepted, this particular form of discrimination could not be easily argued to fall within the scope of the Gender Goods and Services Directive. If an advertisement is purposefully targeted to a vulnerable population it is not denying or curtailing access to a certain service.

Additionally, predatory advertisements are generally targeted towards individuals who pertain to lower socioeconomic classes and socioeconomic status is not a protected ground included within the scope of any of the Equality Directives. This problem might be circumvented in some cases. Low socioeconomic status tends to correlate with racial/ethnic minority group membership. Race or ethnic group will therefore be generally used as a proxy for socioeconomic status and when a particular individual happens to belong to an ethnic minority and poorer class she will be able to invoke the Race Equality Directive. However, there will be other instances in which individuals will only fall within the category of lower socioeconomic class, hence not being protected by any of the Directives. This specific shortcoming, which does not only affect this form of algorithmic discrimination, will be addressed later on in this chapter. In any case, it is much more likely that predatory advertising can be attacked through the mechanisms that already exist in consumer protection law.[910]

The last type of algorithmic discrimination that can be found in advertising is through the reproduction of negative stereotypes that are associated with protected groups. This type of

[908] ELLIS, E. & WATSON, P., *EU Anti-discrimination Law*, *cit.*, 2012, p. 368; CARACCIOLO DI TORELLA, E., "The principle of gender equality, the goods and services directive and insurance: A conceptual analysis", *Maastricht Journal of European and Comparative Law*, vol. 13, No. 3, p. 343; WACHTER, S., "Affinity profiling and discrimination by association in online behavioural advertising", *cit.*, 2020 (forthcoming), pp. 29-30.
[909] US SENATE COMMITTEE ON COMMERCE, SCIENCE & TRANSPORTATION, MAJORITY STAFF, "A review of the data broker industry…", *cit.*, 2013, p. i.
[910] For instance, Directive 2006/114/EC of the European Parliament and of the Council of 12 December 2006 concerning misleading and comparative advertising aims to combat misleading advertising, which it defines as "any advertising which in any way, including its presentation, deceives or is likely to deceive the persons to whom it is addressed or whom it reaches and which, by reason of its deceptive nature, is likely to affect their economic behaviour or which, for those reasons, injures or is likely to injure a competitor" (article 2.b).

discriminatory action does not impact individuals directly, but helps to perpetuate the stereotypes and values through which vulnerable groups have been traditionally oppressed. While not strictly advertising, search engines, in particular Google, should also be included under this category due to the fact that they do, to a certain extent, act as ad providers when they offer search results and, especially because they are one of the leading (if not the main) sources of information for most individuals.

Hence, the way in which a search engine yields results for certain combinations of words, will shape world views on said elements.[911] If Google searches for the word "lesbian", mainly produce porn-related results, this will heavily contribute to the way in which the collective imaginary views and thinks of homosexual women.[912] Once again, the only possibility of addressing these cases of algorithmic discrimination would be when the negative stereotypes reproduced referred to race but not to gender or any other disadvantaged group.

1.1.3. Protected categories

As it was already indicated, the Directives that protect the right to equality in the access to goods and services only cover race and gender. The Directives on equality and non-discrimination in employment protect against discrimination in employment on the basis of gender and race but also on the basis of religion or belief, disability, age or sexual orientation. This brings about the second shortcoming in the protection against discrimination, as religion, belief, disability, age and sexual orientation are not covered by anti-discrimination legislation on access to goods and services and other elements, such as sexual identity, which lead to the discrimination of non cis-normative individuals, are not protected either in employment or access to goods and services. Hence, using automated tools in order to create profiles that help detect an individual's sexual orientation and then using said information to deny homosexual individuals access to a certain good or service would not fall under the more specific scope of protection offered by EU secondary law.[913]

[911] NOBLE, S. U., *Algorithms of Oppression…, cit.*, 2018.
[912] CADOT, J., "Lesbienne : Google a enfin modifié son algorithme", *Numerama*, 18th July 2019. Available on 3rd March 2020 at: https://www.numerama.com/
[913] WACHTER, S., "Affinity profiling and discrimination by association in online behavioural advertising", *cit.*, 2020 (forthcoming), p. 30.

However, perhaps the most problematic issue is the general disregard that the European anti-discrimination Directives have for discrimination based on socioeconomic status. Both the European Convention on Human Rights (art. 14) and the Charter of Fundamental Rights of the European Union (art. 21) include the prohibition of discrimination on the basis of property and social origin. In addition, article 9 of the Treaty on the Functioning of the European Union establishes an obligation consider the requirement to fight social exclusion when developing European policies. However, neither socioeconomic status nor any of its manifestations, such as property, social origin or social condition, are included in the European Equality Directives.

Moreover, "socioeconomic status" or "poverty" are not protected specifically in any of the cited legal documents, which is highly indicative of the lack of importance that the "social class" intersection has in the construction of equality and non-discrimination in Europe. Even if said characteristics can be inferred from property and social origin or even birth, these are wider categories that do not exclusively or explicitly recognise the discrimination that results by the existence of structures of oppression that systematically disadvantage poorer groups and individuals. Social origin, property and birth do include social class,[914] however, by linking social class to other categories such as birth in a certain ideological community or even being adopted, the significance and relevance of the oppression and discrimination undergone by members of lower socioeconomic strata is downplayed and underrated.

Some countries have taken significant steps to establish anti-discrimination legislation that specifically protects members of lower social classes. For example, in 2016 France passed a law prohibiting discrimination on the grounds of social precariousness.[915] Other countries that include protection against discrimination on the grounds of socioeconomic status are Belgium, the Czech Republic, Italy and Spain.[916] It is also particularly relevant to cite a case of indirect discrimination on the basis of socioeconomic status that was successfully brought before a Belgian Court. In said case, it was found that a landlord who stated in

[914] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", *cit.*, 2018, p. 218.
[915] Act No. 2016-832, 24th June 2016, aimed towards fighting discrimination on the basis of social precariousness.
[916] PINET, J. P. & REDEGELD, T., "Poverty discrimination in Europe", 2018. Available on 28th March 2020 at: https://blogs.atd-quartmonde.org

advertisements for his properties that tenants were required to have open-ended employment contracts incurred in a discriminatory conduct on the basis of both wealth and disability.[917]

The aforementioned Judgment is particularly relevant for there is an evident lack of case law concerning instances of discrimination on the basis of socioeconomic status. Very few cases have been brought before the ECHR and CJEU on the basis of social origin and property and the cases that the indicated courts have examined have not referred to socioeconomic status or any of the many discriminatory situations that members of lower social classes endure.[918] Moreover, since discrimination based on socioeconomic status in many cases takes place with regard to financial or insurance services, the fact that their access to these services will be hindered due to their financial status is accepted, not even considering whether this increased difficulty in comparison to other groups is proportionate or not.[919]

Finally, another problematic issue with bringing forward cases on discrimination based on socioeconomic status under the categories of discrimination based on property or social origin are the difficulties in determining or delimiting membership to the discriminated group in cases of indirectly discriminatory practices. Other specially protected categories such as gender or race are fairly easily delimited. That is not to say that there are no issues since, for example, the discussion on whether self-designated transgender women, should be considered women to all effects is in the origin of one of the most significant fractures within feminist scholarship and the feminist movement.[920] However, social origin is much more complicated to establish, for the lines that mark the divide between different socioeconomic groups are not as clearly cut.[921] This does not mean that this type of discrimination is impossible to prove. For example, if a delivery service is not offered in the postal code areas that have the lowest average rent per capita, there is an obvious discrimination based on socioeconomic status.[922]

---

[917] Court of 1st Instance of Namur, 5th May 2015, Le centre interfédéral pour l'égalité des chances et la lutte contre le racisme et les discriminations v. M. Cristophe.
[918] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", *cit.*, 2018, pp. 216-218.
[919] SCHREURS, W. *et al.*, "Cogitas, ergo sum…", 2008, p. 260.
[920] ORTEGA ARJONILLA, E. & PLATERO MÉNDEZ, R. E., "Movimientos feministas y trans* en la encrucijada: aprendizajes mutuos y conflictos productivos", *Quaderns de Psicologia*, vol. 17, No. 3, 2015, pp. 17-30.
[921] SCHREURS, W. *et al.*, "Cogitas, ergo sum…", *cit.*, 2008, p. 260
[922] INGOLD, D. & SOPER, S., "Amazon doesn't consider the race of its customers. Should it?", *Bloomberg*, 21st April 2016.

The lack of effectiveness of prohibitions to discriminate on the basis of socioeconomic status is particularly worrying if we consider that the increased use of algorithms largely affects poorer groups of the population. Many public aid services are increasingly being managed through automated systems[923] that, although generally do not result in a direct discriminatory impact, they no doubt help to perpetuate the disadvantage and oppression suffered by those groups, whether it is by perpetuating the stigma suffer by welfare recipients through automated welfare fraud detection systems,[924] by making mistakes that individuals with less resources have more difficulties in fixing,[925] by setting up huge databases that include members of said groups, who are therefore subjected to a greater scrutiny than other individuals,[926] or by increasing the social exclusion suffered by those who do not have access to new technologies.[927] Furthermore, certain practices such as offering predatory loans are developed by taking advantage of members of groups that are in particularly precarious situations. Current computing capacity enhances the possibility of only targeting individuals that are the specific prey for these types of products and services,[928] which does not only increase the chances that vulnerable individuals will fall for said offers but also makes them much more difficult to detect.

1.2. PURPOSES FOR WHICH ALGORITHMS ARE USED

Although the Race Equality Directive covers advertising, there are many other forms of stereotyping that perpetuate narratives of domination and that are not prohibited under the EU Equality Directives. Search engine results that sexualise women, the fact that e-assistants are mostly female or that image tagging systems mistake Black people for gorillas help to reinforce stereotypes that respond to traditional narratives of oppression.[929] While ALLEN and MASTERS analyse these particular forms of algorithmic discrimination as harassment, it is doubtful that courts will accept said claims.[930]

[923] EUBANKS, V., *Automating Inequality…*, *cit.*, 2017.
[924] RANCHORDÁS, S. & SCHUURMANS, Y., "Outsourcing the welfare state…", *cit.*, 2020, pp. 1-47.
[925] DE LA CUEVA, J., "El derecho a no ser gobernados mediante algoritmos secretos", *cit.*, 2019. *El Notario del Siglo XXI*, No. 87, 2019. Available on 18th February 2020 at: https://www.elnotario.es/
[926] EUBANKS, V., *Automating Inequality…*, *cit.*, 2017.
[927] LERMAN, J., "Big data and its exclusions", *cit.*, 2013, pp. 55-63.
[928] US SENATE COMMITTEE ON COMMERCE, SCIENCE & TRANSPORTATION, MAJORITY STAFF, "A review of the data broker industry…", *cit.*, 2013.
[929] ALLEN, R. & MASTERS, D., "Artificial Intelligence…", *cit.*, 2020, p. 593.
[930] *Ibidem.*

Additionally, it was previously stated, the increasingly extended use of automated systems to select public aid recipients, places vulnerable populations in a situation in which their lives are heavily controlled by authorities and private corporations and their autonomy and freedom can be severely curtailed. Moreover, as it has already been indicated several times, when algorithms used in social welfare programmes yield erroneous results, said errors cannot be contested as discrimination claims within the framework of special protection offered to disadvantaged groups. This is logical given that these decisions could be considered discriminatory or unfair in a general sense but not discriminatory as resulting from decisions based on protected attributes. However, it is also necessary to consider that those affected by errors yielded by algorithms used in social welfare programmes are generally especially vulnerable individuals and groups that have more difficulty in accessing redress mechanisms. Hence, systems that specifically enable claims regarding social welfare programmes should be developed.

2. THE FORMALISTIC-INDIVIDUALISTIC APPROACH TO DISCRIMINATION

An essential part of the type of discrimination analysed in this dissertation, that is, the one that follows from the historical subordination of certain social or identity-groups, is its collective dimension.[931] Individual instances of discrimination result from the existence of an underlying social structure that has historically placed and still places members of certain groups in a position of disadvantage.[932] However, the right to equality and non-discrimination has been built and structured within a legal tradition that is mainly based on (and used to) individual legal categories and builds the notion of equality around the idea of liberal formal equality.[933] In addition, European anti-discrimination legislation follows from and is based on the US' anti-discriminatory regulatory framework which has heavily influenced the construction of the right to equality and non-discrimination from an individualistic perspective that, in many cases, fails to acknowledge the group or collective dimension of this phenomenon.[934]

The implementation of equality from a mainly individualistic perspective leads to a series of shortcomings in addressing instances of discrimination and, especially, in promoting equality.

[931] BARRÈRE, M. A., "Problemas del derecho antidiscriminatorio…", *cit.*, 2003a, p. 3.
[932] AÑÓN ROIG, M. J., "Principio antidiscriminatorio y determinación de la desventaja", *cit.*, 2013b, pp. 134-135.
[933] *Idem*, p. 135.
[934] BARRÈRE UNZUETA, M. A. & MORONDO TARAMUNDI, D., "Subordiscriminación y discriminación interseccional…", *cit.*, 2011, p. 16.

One of the implications of this individual rights-based approach to discrimination mostly built from the notion of formal equality is the formalistic construction of said rights, by mainly articulating it through the direct/indirect discrimination dichotomy, which in the US translates to the disparate treatment/impact clauses.[935] The combination of these two aspects that vertebrate the construction of anti-discrimination law lead to a series of shortcomings and limitations to the effectiveness of the equality and non-discrimination framework.

2.1. ESTABLISHING THE DIFFERENCE BETWEEN DIRECT AND INDIRECT DISCRIMINATION

One of the first issues that arises with regard to the direct/indirect discrimination dichotomy is that, in some instances, establishing the difference between direct and indirect discrimination is very difficult in practice.[936] For example, when discrimination occurs with regard to elements that are inseparable from the protected ground (for instance, if a regulation only allows individuals who had been married to their partners to access widowers pension-schemes in a country in which homosexuals are not allowed to get married),[937] the consideration of such a situation as a case of direct or indirect discrimination will largely depend on the interpretation carried out by the Court. Determining whether an apparently neutral provision affects the entire population of an oppressed group can be very important when differentiating between cases of direct and indirect discrimination. In this sense, the CJEU has considered that when a formal criterion affects the whole protected group it is constitutive of a situation of direct discrimination.[938]

However, and while in some cases, situations of discrimination that are not clearly cut as either direct or indirect discrimination are ruled by courts as instances of the former, the criteria for considering that a case falls within the scope of application of the direct discrimination prohibition is generally very strict, meaning that courts generally choose to apply the indirect discrimination category.[939] The implications this has for algorithmic discrimination is that, unless full transparency and understandability of automated or semi-automated decisions is granted, thereby providing the tools to detect instances of direct discrimination, most cases will have to be redirected through the indirect discrimination

[935] *Ibidem*.
[936] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", *cit.*, 2018, p. 54.
[937] CJEU Judgment 1st April 2008, C-267/06, Tadao Maruko v. Versorgungsanstalt der deutschen Bühnen.
[938] *Ibidem*.
[939] XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *Op. cit*., 2020, pp. 172-173.

pathway which offers greater possibilities for defendants to justify their actions and therefore limits the effectiveness of the anti-discrimination framework.[940]

2.2. INTERSECTIONAL DISCRIMINATION

Another result of articulating the anti-discriminatory framework through individual rights and the direct/indirect discrimination dichotomy which, does, to a certain extent, accept and acknowledge the demands of emancipatory theories and political movements such as feminism and anti-racism, is the creation of specific and static protected categories of discrimination, in other words, failing to acknowledge the existence of intersectional discrimination.[941]

As was already indicated in chapter II, using intersectional perspectives is not always useful when researching and analysing discrimination in general, due to the wide variety of disadvantaged groups that a single individual can belong to. However, it is possible to address discrimination from intersectional approaches on a case-by-case basis, which is the reason why the possibility of basing lawsuits on intersectional discrimination must be included in anti-discrimination regulatory instruments.

Although intersectional (or multiple) discrimination is not comprehensively regulated, it is mentioned in Recital 14 of the Racial Equality Directive and Recital 3 of the Employment Equality Directive, both of which state that "women are often the victims of multiple discrimination". However, the formalistic approach to discrimination both through the direct-indirect dichotomy and the static categorisation of specially protected groups has led the CJEU to reject the possibility of ruling certain rules or actions as discriminatory if the grounds for discrimination resulted from the intersection of two different protected groups. This is due to the fact that the grounds for discrimination contained in EU secondary law are a *numerus clausus*, which means that the CJEU cannot establish new grounds for discrimination unless it chose to carry out an extensive interpretation that integrated the anti-discrimination Directives with articles 20 and 21 of the Charter of Fundamental Rights of the EU which do contain an open clause regarding discrimination grounds.[942]

---

[940] *Ibidem*.
[941] BARRÈRE UNZUETA, M. A. & MORONDO TARAMUNDI, D., "Subordiscriminación y discriminación interseccional…", *cit.*, 2011, pp. 29-30.
[942] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", *cit.*, 2018, p. 62.

Consequently, although discrimination can be based on more than one ground included in the Directive that may be of application to the specific case, several grounds cannot be combined in order to create a new subgroup susceptible of being discriminated with its own particularities and specific experiences of discrimination such as "black women".[943] For example, in *Parris v. Trinity College and others*[944] the CJEU ruled that:

> "Articles 2 and 6(2) of Directive 2000/78 must be interpreted as meaning that a national rule such as that at issue in the main proceedings is not capable of creating discrimination as a result of the combined effect of sexual orientation and age, where that rule does not constitute discrimination either on the ground of sexual orientation or on the ground of age taken in isolation".[945]

Additionally, some European states have specifically addressed intersectional discrimination in their regulatory instruments. For example, section 4 of the German General Act on Equal Treatment covers "unequal treatment on several grounds" although only with regard to the possible justifications to this specific kind of unequal treatment.

Moreover, the intersectional discrimination perspective has also been introduced in the regulation of affirmative action in Bulgaria seeing as article 11 of the Bulgarian Protection Against Discrimination Act indicates that "the bodies of state power, the public bodies and the local government bodies shall take priority measures […] to equalise the opportunities of persons who are victims of multiple discrimination".

It is also interesting to highlight the way in which the Croatian Anti-discrimination Act approaches intersectional discrimination. Article 6 of said regulatory instrument establishes four types of discrimination which are labelled as "more serious" and which consist of multiple discrimination, "discrimination committed several times (repeated discrimination), discrimination which lasted a longer period of time (continued discrimination), or discrimination whose consequences are particularly harmful for the victim". Courts must take

[943] *Idem*, p. 63.
[944] *Ibidem*: "In *Parris v. Trinity College and Others*, the CJEU had to address the possibility of multiple discrimination, since the referring court specifically posed this question. Dr Parris requested that on his death the survivor's pension provided for by the pension scheme should be granted to his civil same-sex partner. He was refused on the basis that they entered into a civil partnership only after he had turned 60, thus not meeting the pension scheme requirements. The civil partnership, however, was established in the United Kingdom in 2009, once Dr Parris was over 60 years old; in Ireland, it was only recognised from 2011 onwards. This meant that any homosexual person born before 1 January 1951 would not be able to claim a survivor's benefit for his civil partner or spouse under this scheme".
[945] CJEU Judgment 24th November 2016, C-443/15, David L. Parris v. Trinity College Dublin and Others.

into consideration these more serious forms of discrimination when establishing immaterial damages.

However, even when there are provisions that specifically address intersectional discrimination, their applicability has been generally scarce and very limited.[946] One of the very few courts that has started to apply anti-discrimination law in a more flexible manner that is not solely reduced to direct and indirect discrimination or to the protected groups as static categories is the European Court of Human Rights (ECHR), which has recognised the specific vulnerabilities suffered individuals who belong to more than one oppressed group. For example, in the case of B.S. v. Spain, the Court stated:

> "In the light of the evidence provided in this case, the Court considers that the decisions issued by the domestic Jurisdictional Bodies did not take into account the specific vulnerability of the applicant, inherent in her condition as an African woman exercising prostitution. The Authorities thus failed to comply with their obligation, under article 14 of the Convention combined with article 3, to take all possible measures to see whether a discriminatory attitude could or could not have played a role in the events."[947]

Furthermore, intersectional discrimination crossing gender and age axes has also been admitted:

> "The question at issue here is not considerations of age or sex as such, but rather the assumption that sexuality is not as important for a fifty-year-old woman and mother of two children as for someone of a younger age. That assumption reflects a traditional idea of female sexuality as being essentially linked to child-*bearing* purposes and thus ignores its physical and psychological relevance for the self-fulfilment of women as people."[948]

While the fact that this perspective is already being applied by the ECHR, intersectional discrimination and the particularities of the identities of individuals who suffer it are still very difficult to prove given the way in which legal systems are shaped in Western societies. The

[946] CHOPIN, I. & GERMAIN, C., "A comparative analysis of non-discrimination law in Europe 2019", *cit.*, 2019, p. 42.
[947] ECHR Judgment24th July 2012, 47159/08, B.S. v. Spain, 47159/08, paragraph 71.
[948] ECHR Judgment 25th July 2017, 17484/15, Judgment Carvalho Pinta v. Portugal, paragraph 52.

very fact that the previously cited cases reached the ECHR means that intersectional perspectives were not being applied by national courts.

2.3. STRUCTURAL OR SYSTEMIC DISCRIMINATION

In many cases, when discrimination is analysed, this analysis is carried out from traditional liberal perspectives. This means that although the specific cases in which an individual or group is discriminated are acknowledged, they are not analysed from the perspective of the deeper social discriminatory structure in which they are embedded. This is precisely the perspective that is adopted by most Western anti-discrimination legal instruments and court decisions, which, by limiting their approach to discrimination to the duality between direct and indirect discrimination, do not sufficiently, address the existence of general structures of social subordination.[949]

The *Bilka-Kaufhaus*[950] case shows the way in which the CJEU fails to introduce structural discrimination in its judgments. In this case, part-time workers were excluded from a series of pension benefits, which are considered as part of workers' payment under the EU framework. This resulted in a much more negative impact on women than men seeing as most part-time workers were (and still are) women. While the CJEU indicated that this constituted an indirectly discriminatory measure it opened the door to its justification on economic grounds and, more importantly, considered that the equal pay provision contained in article 157 of the TFEU (then article 119 of the Treaty establishing the European Economic Community) did not require employers "to organize [their] occupational pension scheme in such a manner as to take into account the particular difficulties faced by persons with family responsibilities in meeting the conditions for entitlement to such a pension".[951]

The Court then failed to acknowledge, or did not give sufficient weight, to the fact that women take part-time jobs at a higher rate than men due to the caring responsibilities that society continues to assign them. More recent judgments, such as the ones delivered in the *Achbita*[952] and *Bougnaoui*[953] cases, also show how European institutions lack the deeper

[949] AÑÓN ROIG, M. J., "Principio antidiscriminatorio y determinación de la desventaja", *cit.*, 2013b, p. 130; BARRÈRE UNZUETA, M. A., "Problemas del derecho antidiscriminatorio…", *cit.*, 2003a, pp. 10-11.
[950] CJEU Judgment 13th May 1986, C-170/84, Bilka - Kaufhaus GmbH v. Karin Weber von Hartz.
[951] *Idem*, paragraph 43.
[952] CJEU Judgment 14th March 2017, C-157/15, Samira Achbita and Centrum voor gelijkheid van kansen en voor racismebestrijding v. G4S Secure Solutions NV.

analysis that discrimination cases require. Moreover, the very restrictive interpretation of provisions regarding the implementation of positive action measures to advance the equality of traditionally disadvantaged groups carried out by the CJEU further proves the Court's failure to grasp and introduce the existence of systemic discrimination in its analysis of the rights to equality and non-discrimination.

In addition, the CJEU's general reluctance to accept what they consider to be measures of "positive discrimination", also conveys how it fails to grasp the existence of a structure of disadvantage that, in some cases, in order to change, needs to operate, at least temporarily, through mechanisms that focus on achieving equality of results.

Finally, as the previous section conveys, the failure to properly address the existence of a systemic disadvantage of certain groups also takes place through the lack of a comprehensive regulation of intersectional discrimination as well courts' reluctance to apply theories of intersectional discrimination when having to address these situations. This is due to the fact that in order to address intersectional discrimination it is necessary to fully analyse the pervasiveness of the social constructions that underlie the specific forms of discrimination undergone by members of more than one disadvantaged group.[954]

2.4. THE NEED FOR A COMPARATOR

With the exception of very specific cases, courts generally demand a comparator in order to establish the existence of unequal treatment or of a *prima facie* case of unequal impact. A salient example of a case in which a comparator is not requested is when women receive a harmful treatment, such as being denied a promotion, because they are pregnant. In these cases, the CJEU considered that, an unfavourable treatment that directly results from a woman being pregnant constitutes a case of direct discrimination on the basis of sex even if the treatment provided to a male counterpart is not provided.[955]

When dealing with instances of algorithmic discrimination it may not be possible to provide a comparator if the algorithm is not transparent and/or if the output data is protected by data protection rights. As it was already indicated, requesting aggregate anonymised data and

[953] CJEU Judgment 14th March 2017, C-188/15, Asma Bougnaoui and Association de défense des droits de l'homme (ADDH) v. Micropole SA.
[954] AÑÓN ROIG, M. J., "Principio antidiscriminatorio y determinación de la desventaja", *cit.*, 2013b, p. 135.
[955] CJEU Judgment 8th November, 1990, C-177/88, Elisabeth Johanna Pacifica Dekker v. Stichting Vormingscentrum voor Jong Volwassenen (VJV Centrum) Plus.

providing plaintiffs with the possibility of experimenting with the algorithm may help to overcome the obstacles specified above.

Another related issue is the ever-evolving nature of some automated systems, which may, in some cases mean, that no two individuals have been subjected to the exact decision-making process, thereby hindering the possibility of providing a comparator. It will therefore be necessary to force system programmers to develop algorithms so that controllers can store every version of the software programme.

Nonetheless, there are still instances in cases of algorithmic and normal discrimination in which it may not be possible to provide a comparator, in which case, the introduction of notions of disadvantage, subordination, structural and intersectional discrimination, as interpretative criteria, are necessary in order to understand the specific nature of harms caused to members of disadvantaged groups.[956]

3. ENFORCEMENT

3.1. GENERAL ENFORCEMENT MECHANISMS

The equality and anti-discrimination framework can be enforced through civil, criminal and administrative procedures.[957] National enforcement systems are mainly focused on remedying instances of discrimination rather than preventing them.[958] One of the key ways in which discrimination could be prevented is through effective oversight equality bodies. However, at the EU level, the European Institute of Gender Equality (EIGE) and the Fundamental Rights Agency (FRA), are the only two bodies which are competent in matters of equality and non-discrimination and the activities they carry out are mainly focused on research and data collection. Additionally, the European Commission can only bring about claims against member states for non-compliance with the rights to equality and non-discrimination but not against private parties, therefore leaving it up to individuals or national equality bodies to bring about discrimination claims against private parties.

[956] AÑÓN ROIG, M. J., "Principio antidiscriminatorio y determinación de la desventaja", *cit.*, 2013b, pp. 146-152.
[957] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", *cit.*, 2018, p. 248.
[958] CHOPIN, I. & GERMAIN, C., "A comparative analysis of non-discrimination law in Europe 2019", *cit.*, 2019, p. 100.

The Directives on Racial and Gender Equality establish that all member states of the EU "designate a body or bodies for the promotion of equal treatment of all persons without discrimination on the [protected] grounds."[959] All member states have thus put in place equality bodies. Most have established bodies that deal with all protected grounds of discrimination, including those in the Employment Equality Directive and others that are not comprehended under any of the EU Equality Directives.[960] Some countries have also designated specific bodies to cover one of the protected grounds, mainly gender or race.[961] However, these bodies lack the oversight capacity that would be necessary for an effective system of prevention of discriminatory activities.

3.2. ACCESS TO JUSTICE

Access to justice can be approached from two perspectives when analysed in combination to the rights to equality and non-discrimination. On the one hand, it generally refers to ensuring that all individuals have the possibility of bringing claims of violations of the rights to equality and non-discrimination before the courts.[962] On the other hand, it can be viewed from a more specific approach as the extent to which victims of discrimination obtain redress and the prohibitions to discriminate are properly enforced.[963]

The rights to an effective remedy and a fair trial are recognised by the European Convention of Human Rights (arts. 6 and 13) and the Charter of Fundamental Rights of the European Union (art. 47). Additionally, the EU Equality Directives establish that "member states shall ensure that judicial and/or administrative procedures … for the enforcement of obligations under this Directive are available to all persons who consider themselves wronged by failure to apply the principle of equal treatment to them…".[964] The CJEU has indicated that said provisions mean "that the member states must take measures which are sufficiently effective

[959] Art. 13 Racial Equality Directive; Art. 20 Gender Employment Equality Directive; and, art. 12 Gender Goods and Services Directive.
[960] BURRI, S., SENDEN, L. & TIMMER, A., "A comparative analysis of gender equality law in Europe 2019", Brussels, European Commission, 2019, p. 142; CHOPIN, I. & GERMAIN, C., "A comparative analysis of non-discrimination law in Europe 2019", *cit.*, 2019, pp. 105-110.
[961] *Ibidem.*
[962] GERARDS, J. & GLAS, L. R., "Access to justice in the European Convention of Human Rights system", *Netherlands Quarterly of Human Rights*, vol. 35, No. 1, 2017, p. 13.
[963] *Ibidem*.
[964] Employment Equality Directive, Art. 9.1; Gender Employment Equality Directive, Art. 17.1; Gender Goods and Services Directive, Art. 8.1; Racial Equality Directive, Art. 7.1.

to achieve the aim of the directive and that they must ensure that the rights thus conferred may be effectively relied upon before the national courts by the persons concerned".[965]

General limitations to access to justice take place as a result of the bureaucratic and economic cost that judicial procedures entail, which automatically places members of disadvantaged groups in worse positions of departure.[966] There are also other procedural barriers, such as very short time frames for bringing cases before the courts, that factor into limiting individuals' chances of effectively accessing justice.[967] Additionally, when it comes to bringing cases before supranational courts, said limitations are further enhanced due to their limited powers of review, especially when it comes to judging private discriminatory practices.[968]

Another constraint on access to justice in discrimination cases, which is further enhanced through the growing use of automated decision-making, is that individuals may not even be aware that they are being discriminated against. This is also the case when humans make decisions regarding, for example, loan eligibility or hiring. In these instances, only if the individual suspects that the decision is unfair in some way and asks for it to be reviewed or if she has knowledge that there are significant differences between the outcomes for members and non-members of the protected group, will she be able to detect the existence of discriminatory practices.

The chances of detecting the existence of discriminatory practices by its victims can be further undermined when automated decision-making takes place due to the intervention of privacy regulations. Data protection regulations provide processors and controllers with the possibility of limiting access to said information. Additionally, identifying the liable party will prove increasingly hard[969] as there are varying levels and models of subcontracting digital and data processing services, algorithms continuously evolve and, in some cases, can be modified by an unidentifiable number of programmers.

[965] CJEU Judgment 15th May 1986, C-222/84, Johnston v. Chief Constable of the Royal Ulster Constabulary, paragraph 17.
[966] CHOPIN, I. & GERMAIN, C., "A comparative analysis of non-discrimination law in Europe 2019", Brussels, European Commission, 2019, pp. 83-84.
[967] *Ibidem*.
[968] XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *cit.*, p. 175.
[969] *Idem*, p. 174.

Finally, it is also necessary to point out an issue that is closely related with the individualistic-formalistic perspectives from which the European equality and anti-discrimination is built but that have very significant implications in the effectiveness of anti-discrimination claims and thus the enforcement of the equality and non-discrimination framework. Although the EU anti-discrimination framework is inspired by the American system, there are two elements that have been crucial for disparate impact (indirect discrimination) claims to succeed and which have, nonetheless, not been developed to the same extent in Europe. These elements are, statistical evidence and class actions.

The use of statistical evidence will, no doubt, become increasingly important as cases of algorithmic discrimination are brought before the courts. Class actions allow for both the individual and collective dimensions of discrimination to be taken into consideration and to protect members of the group that may not have the resources or willpower to fight discrimination cases,[970] but are not allowed under the European equality framework. While some member states have introduced some form of collective redress either through class actions or *actio popularis*,[971] that is, the possibility that associations and organisations have in bringing cases that affect the interests they represent, these mechanisms have not gained a great degree of traction in discrimination cases.[972]

3.3. INSTRUCTIONS TO DISCRIMINATE

"Instruction to discriminate" is a type of discriminatory attitude comprehended under the EU Equality Directives[973] and most member states' equality frameworks.[974] However, none of the Directives describes what giving instructions to discriminate must entail in order to be considered itself a discriminatory attitude and has therefore not had much practical application.[975] The reason why this particular type of discriminatory behaviour is included under the enforcement category is that, if properly articulated, it could help to render

[970] BARNARD, C. & HEPPLE, B., "Indirect discrimination: interpreting Seymour-Smith", *Cambridge Law Journal*, vol. 58, No. 2, 1999, pp. 401-402.
[971] CHOPIN, I. & GERMAIN, C., "A comparative analysis of non-discrimination law in Europe 2019", *cit.*, 2019, pp. 92-94.
[972] NAGY, C. I., *Collective Actions in Europe: A Comparative, Economic and Transsystemic Analysis*, Cham, Springer, 2019, pp. 73-85.
[973] Art. 2.4 Employment Equality Directive; art. 2.2.b Gender Employment Equality Directive; art. 2.4 Racial Equality Directive and art. 4.4 Gender Goods and Services Directive.
[974] CHOPIN, I. & GERMAIN, C., "A comparative analysis of non-discrimination law in Europe 2019", *cit.*, 2019, pp. 50-51.
[975] EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", *cit.*, 2018, pp. 64-69.

programmers, processors and controllers responsible for the discriminatory results of algorithms and thus act as an enforcement mechanism.

In this sense, it is relevant to highlight that this is yet another of the many mechanisms that the equality and non-discrimination European legal frameworks offers and that can be adapted in order to control instances of algorithmic discrimination. Many of the shortcomings that have been detected are the result of the way in which the current framework is applied and interpreted but the existing basis does provide a series of very useful tools in order to deal with algorithmic discrimination. However, for the full potential of these mechanisms to be effectively implemented in the control of instances of discrimination that result from the use of automated systems, it is necessary to develop specific protocols for courts and public bodies in general. In particular, with regard to the judicial review of algorithmic discrimination cases, there are certain elements such as rules for presenting proof and expert witnesses that may have to undergo significant modifications in order to adapt to the needs of these specific forms of discrimination. This necessary shift in certain judicial traditions and in the control and protection of equality and non-discrimination may also, in turn, prove useful in addressing the general shortcomings and limitations that the current framework for the protection of the rights to equality and non-discrimination and its implementation currently have.

However, due to the specific nature of algorithmic discrimination, that is, the form of discrimination and perpetuation of inequality that is mediated by automated systems, it is necessary to complement the application of equality and non-discrimination rules with other regulatory instruments that specifically and comprehensively address algorithms and the risks and harms they generate. The following part sets out to address the data protection framework as the main existing regulatory framework that is currently being used to tackle the general risks and harms caused by automated decision-making and briefly establishes a series of proposals aimed towards addressing the insufficiencies of existing regulatory instruments in the protection against the risks generated by algorithms and, in particular, algorithmic discrimination.

PART II. REGULATING ALGORITHMS

There are many legal implications that result from the widespread use of algorithms. This dissertation focuses on the ways in which algorithms impact individuals' fundamental rights to equality and non-discrimination. As the previous part shows, only focusing on controlling algorithmic discrimination from the perspective of anti-discrimination law is insufficient. These insufficiencies partly result from the fact that the current equality and anti-discrimination law framework falls short in the protection it offers individuals and groups against all instances of discrimination, regardless of whether these are the result of exclusively human decisions or of decisions mediated by machines.

However, the shortcomings presented when aiming to control algorithmic discrimination from the perspective of the legal framework of protection to the rights of equality and non-discrimination also stem from the specific nature and risks produced by algorithms, which were already briefly addressed in the first chapter in part one.

As it has already been indicated several times, algorithmic discrimination is a double-pronged phenomenon. On the one hand we find discrimination, which results from social norms and institutions being built by members of those groups that have historically held power. In this sense, algorithms operate yet as another construction that allows for the perpetuation of power structures that disadvantage the members of certain groups. From this perspective, algorithmic discrimination must be subsumed in the equality and non-discrimination legal framework. On the other hand, the second prong to the phenomenon of algorithmic discrimination are algorithms and the novel social and regulatory challenges they pose. Algorithms produce a series of problems that are all closely linked between each other, amongst which we find discrimination.

Hence, algorithmic discrimination can and should be protected by the legal framework focused on protecting the rights to equality and non-discrimination but it should also be addressed from the perspective of a public regulatory framework that aims to control and solve all the different issues generated by algorithms. In order to effectively control algorithmic discrimination, it is therefore necessary to comprehensively address many of the issues that arise with regard to the use of algorithms and the impact they can have on the rights of individuals and on society as a whole.

With this in mind, this second part to the dissertation is structured as follows: chapter I explains the different concerns that arise from the use of data processing technologies and briefly refers to the way in which said concerns have thus far been addressed by the data protection framework in Europe; chapter II analyses the data protection framework in Europe and its applicability to the risks generated by automated systems; chapter III covers the shortcomings of the privacy perspective and briefly addresses the possibility and need of combining the existing anti-discrimination and data protection European legal frameworks in order to deal with algorithmic discrimination. The final chapter puts forward a series of proposals

CHAPTER I. ALGORITHMIC RISKS AND HARMS AND THE EU DATA PROTECTION SOLUTION

1. GENERAL CONCERNS REGARDING THE USE OF AUTOMATED SYSTEMS

1.1. UNFAIR (AND DISCRIMINATORY) OUTCOMES

1.1.1. Biased humans and accurate machines

As KAHNEMAN and TVERSKY showed, human-made decisions are influenced by a series of mental shortcuts that individuals take when judging or analysing situations.[976] These shortcuts are named "heuristics".[977] For example, by using the representativeness heuristic, people will assess that an individual described as shy is far more likely to be a librarian than a salesperson since her description aligns with stereotypes held for the former far more so than for the latter.[978] In other words, because the attribute shy represents the stereotype assigned to librarians, through the use of the representativeness heuristic, people described as shy will be thought to be librarians at a much higher rate than other professions. While the use of these shortcuts or heuristics can be effective, they lead to decisions based on partial information and can therefore generate cognitive biases.[979]

Automated decision-making tools are fed with all the information that has to be processed and instructions for processing. Hence, machines tend to offer more objective and precise results than human decision-makers.[980] The use of algorithms makes all types of processes increasingly efficient, from more mechanical or scientific types of prediction such as establishing the level of toxicity in chemical compounds[981] to predictive systems that affect individuals by, for instance, creating personal profiles that help firms target product advertisements to certain audiences.[982]

---

[976] TVERSKY, A. & KAHNEMAN, D., "Judgment under uncertainty…", *cit.*, 1974, pp. 1124-1131.
[977] *Ibidem.*
[978] *Idem*, p. 1124.
[979] *Idem*, p. 1131.
[980] CHOULDENOVA, A., "Fair prediction with disparate impact: a study of bias in recidivism prediction instruments", 2016, p. 5. Available on 9th April 2019 at: https://arxiv.org/; MILLER, A. P., "Want less-biased decisions? Use algorithms", *Harvard Business Review*, 26th July 2018. Available on 13th February 2019 at: https://hbr.org/
[981] COGLIANESE, C. & LEHR, D., "Regulating by robot…" *cit.*, 2017, p. 1162.
[982] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, p. 70.

Since algorithms generally produce more accurate and efficient decisions than humans, it could be argued that there is no need to develop new regulatory instruments, for existing regulations should be able to address a type of decision-making that is overall more precise than human decision-making.[983] In theory, existing regulatory instruments should provide individuals with the necessary mechanisms to defend themselves when they consider that a decision based on automated processing has treated them unfairly or has led to inaccurate results that affect them in some way.

1.1.2. Measuring baseball vs. measuring humans

In general, machine learning algorithms can provide fair and objective results when the purpose for which they are deployed is easily measurable. For example, an area in which the use algorithms has been particularly successful is in baseball.[984] Additionally, the fact that the use of algorithms in predictive systems used, for instance, in sports or GPS tools such as Google Maps, does not affect individuals' rights means that the need for regulation is not as pressing.[985] Moreover, even when algorithms are used in processes that affect individuals, as long as the problems specified are straightforward, they still achieve a significant degree of objectivity and accuracy.[986]

However, in many of the cases in which automated or semi-automated processes affect individuals, the difficulty of specifying real-life problems (such as creditworthiness) as mathematical measurements[987] can easily lead to erroneous, biased or discriminatory outcomes. Moreover, as many scholars have stated, the way in which algorithms are shaped and the logics they follow are generally mediated by the personal values and prejudices of their designers.[988] The origin of these biases, which are then consciously or unconsciously embedded into algorithmic systems, generally lies in the personal experiences of system designers and are reinforced by the social structures that surround them and in which algorithms are generated. These social structures are both the more specific environments in

[983] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1538-1540.
[984] O' NEIL, *Weapons of Math Destruction*, *cit.*, pp. 15-17.
[985] KOCHI, E., "How to prevent discriminatory outcomes in machine learning", *cit.*, 2018, p. 10.
[986] KLEINBERG, J. *et al.*, "Discrimination in the age of algorithms", *cit.*, 2018, p. 134.
[987] LEHR, D. & OHM, P., "Playing with the data…" *cit.*, 2017, pp. 672-673.
[988] MACNISH, K., "Unblinking eyes: the ethics of automating surveillance", *Ethics and Information Technology*, vol. 14, No. 2, 2012, p. 158.

which automated systems are created, such as big tech companies, as well as the social values and constructs present in the society in which algorithms are created.[989]

1.1.3. Human bias and machine error

The high degree of human intervention in the design and deployment of algorithms may not only lead to reinforcing certain stereotypes and pre-existing situations of oppression by yielding discriminatory results. While algorithmic discrimination is the main focus of the dissertation, there are also other types of flaws that humans introduce in automated systems that can also harm individuals subjected to automated or semi-automated decision-making. There are many instances in which the very same errors that humans make when using mental shortcuts can also be made in the process of constructing the algorithm,[990] leading to erroneous and/or biased results.

Biases or errors in the final decision can be developed as a result of biased, erroneous and unrepresentative datasets[991] or as a consequence of "technological constraints, errors or design decisions",[992] such as flawed randomisation commands.[993] Error and bias can also be introduced as a consequence of the self-learning abilities of automated systems. It is not uncommon to find examples in which algorithms that are initially neutral end up yielding discriminatory results or reinforcing narratives of oppression as it learns from the actions of its users.[994] Self-development can also lead to problematic outcomes that, although not discriminatory in nature, can harm individuals subjected to automated processing. If the system is not properly taught how to adapt to developments in the society it is deployed, mismatches between the algorithmic system and the reality it is set to function in may arise.[995]

It is also possible that algorithms reach erroneous conclusions due to their inability to integrate and consider certain elements that a human decision-maker would be able to take

[989] FRIEDMAN, B. & NISSEMBAUM, H., "Bias in computer systems", *ACM Transactions on Information Systems*, vol. 14, No. 3, 1996, p. 334.
[990] KAMINSKI, M. E., "Binary governance…", *cit*., 2019b, pp. 1538-1539.
[991] CUKIER, K., and MAYER-SCHOENBERGER, V., "The rise of big data…", *cit,* 2013, p. 29; BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016,, p. 681; KIM, P. T., "Data-driven discrimination at work", *cit.*, 2017, pp. 885-886.
[992] MITTELSTADT, B. D. *et al.*, "The ethics of algorithms…", *cit.*, 2016, p. 7.
[993] FRIEDMAN, B. & NISSEMBAUM, H., "Bias in computer systems", *cit.*, 1996, p. 334.
[994] SWEENEY, L., "Discrimination in online ad delivery", *cit.*, 2013, p. 52; KOCHI, E., "How to prevent discriminatory outcomes in machine learning", *cit.*, 2018, p. 10.
[995] MITTELSTADT, B. D. *et al.*, "The ethics of algorithms…", *cit.*, 2016, p. 8; FRIEDMAN, B. & NISSEMBAUM, H., "Bias in computer systems", *cit.*, 1996, p. 335.

into account. This results from the fact that it is not always possible to predict and teach an automated system all the possible behaviours that the individuals subjected to algorithmic decision-making might adopt which results in these systems omitting certain contextual knowledge that could be relevant for the decision.

For example, the algorithms used by school districts in the US in order to evaluate teachers and fire those who scored the lowest has been heavily criticised and covered by the scholarship due to the fact that some of the teachers fired were considered to be very good by parents, students and their superiors. Since the models mostly rely on test scores and compare the evolution of each class from year to year, it is suspected that some teachers correct their students' responses in order to increase their own teaching score.[996] These are elements that the algorithm does not necessarily capture and correct for, which leads to unfair decisions being made regarding those teachers who correctly follow the system's rules but who get a class whose grades were artificially inflated by the previous teacher.[997]

It is thus relevant to keep in mind that although human decision-making may be fundamentally biased, humans have the possibility of adapting their views to the context of each case and observe certain intangible elements that algorithms may not detect. Moreover, although human biases have been extensively classified and documented,[998] humans do not always make the same mistakes or analyse situations in the same manner. However, when an individual designs an algorithm and, either conscious or unconsciously, embeds a series of biases in it, the automated system will always be constrained by said biased parameters. These systems currently still lack the capacity that humans have of producing a real case-by-case analysis and will therefore always produce its results based on the same bias.

In addition, algorithmic decision-making does not establish causality but reaches conclusions based on correlations.[999] The large volumes of data that are currently available and the computing capacity for processing said information means that decisions made by algorithms based on correlations are generally considered to be reliable enough without establishing causality.[1000] The final results obtained from the data are therefore based on correlations but

[996] O' NEIL, *Weapons of Math Destruction*, *cit.*, pp. 3-11; KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1531.
[997] *Ibidem.*
[998] TVERSKY, A. & KAHNEMAN, D., "Judgment under uncertainty…", *cit.*, 1974, pp. 1124-1131.
[999] MITTELSTADT, B. D. *et al.*, "The ethics of algorithms…", *cit.*, 2016, p. 5; FRIEDMAN, B. & NISSEMBAUM, H., "Bias in computer systems", *cit.*, 1996, p. 334.
[1000] CUKIER, K., and MAYER-SCHOENBERGER, V., "The rise of big data…", *cit.*, 2013, p. 29

treated as causal relationships, generating a risk of faulty inferences.[1001] LOHR illustrates this problem quite clearly in the following way:

> "Imagine spending a few hours looking online for information on deep fat fryers. You could be looking for a gift for a friend or researching a report for cooking school. But to a data miner, tracking your online viewing, this hunt could be read as a telltale sign of an unhealthy habit — a data-based prediction that could make its way to a health insurer or potential employer".[1002]

It is highly likely that, in some cases, what is true for a certain group in general, does not apply to a particular case. Taking the example indicated above, the fact that most individuals who search for deep fat fryers online are more likely of incurring in unhealthy habits might be accurate. However, if this general notion is actually not applicable in a particular case, such as the one stated above, seriously negative, unfair and unjustified consequences will be delivered for the individual subjected to automated decision-making.

All in all, there are many opportunities during the development and after the deployment of algorithms for these systems to incorporate bias, prejudice and errors that will lead to skewed and/or erroneous results or to the reinforcement of pre-existing situations of oppression.[1003] After all, automated systems are mere simplifications of reality and are therefore sometimes unable to capture all of the specificities that intervene in human behaviour, decisions and interactions.[1004]

1.1.4. The technological heuristic

While, in some cases, algorithms have shown to correct for existing biases and errors in the data on their own,[1005] this has happened when the other elements in the algorithm that could generate skewed results have been controlled and shaped in detail to correct for said biases.[1006] Hence, seeing as the individuals and teams in charge of building algorithms will

[1001] BAROCAS, S., "Data mining and the discourse on discrimination", *cit.*, 2014, p. 2.
[1002] LOHR, S., "Sizing up big data, broadening beyond the Internet", *The New York Times*, 29th June 2013. Available on 28th April 2019 at: https://bits.blogs.nytimes.com/
[1003] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1540.
[1004] BAROCAS, S., "Data mining and the discourse on discrimination", *cit.*, 2014, p. 2.
[1005] EREL, I. *et al.*, "Selecting directors using machine learning" *Fisher College of Business Working Paper No. 2018-03-005*; *European Corporate Governance Institute (ECGI) - Finance Working Paper No. 605/2019*, 2019, pp. 20, 31. Available on 7th July 2019 at: https://ssrn.com/
[1006] KLEINBERG, J. *et al*., "Human decisions and machine predictions", *cit.*, 2017, pp. 244-245.

not always be aware of the risks of these tools, it is highly likely that some of the choices made in the process of building the algorithm lead to erroneous or biased outcomes.[1007]

An additional problem arises due to the fact that people tend to trust automated systems as impartial and objective decision-makers.[1008] This can, in turn, generate a common agreement that regulating algorithmic decision-making is not necessary. Furthermore, there is a general tendency to accept the results that automated systems provide as neutral and valid,[1009] increasing the chances that said results are taken at face value and not challenged or questioned.[1010] This excessive reliance on automated systems and the accuracy and validity of the results they produce can be analysed as a new form of heuristic or mental shortcut: a technological heuristic.[1011]

However, this must not be taken to mean that including a "human-in-the-loop" is not desirable. For instance, in a study regarding an algorithm used to assist child welfare services, DE-ARTEAGA *et al.* show that humans did, in many cases, detect and override recommendations made by the software programme based on erroneous child risk estimates.[1012] In the study they conducted, the "humans in the loop" were specialised social workers who had access to all the information relative to each particular case. Hence, while the technological heuristic is a risk that must be considered, if properly set up, "humans in the loop" can be an essential mechanism for effective algorithmic oversight.

1.1.5. Regulating algorithms to prevent and deal with biases, errors and discrimination

All the possibilities for algorithmic errors that have been indicated must not be taken mean that the use of automated systems should be discarded altogether, but that great care should be taken when implementing algorithmic decision-making, particularly when they can impact the fundamental rights of individuals. An excessive reliance on the supposed objectivity and accuracy of algorithms generates very significant risks for human rights due to the fact that these automated tools, while generally more precise than human decision-makers, can also

[1007] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, pp. 145-146.
[1008] CITRON, D. K., "Technological due process", *cit.*, 2008, pp. 1271-1272.
[1009] CITRON, D. K., "Technological due process", *cit.*, 2008, pp. 1271-1272; PARASURAMAN, R. & MILLER, C. A., "Trust and etiquette in high-criticality automated systems", *cit.*, 2004, p. 52.
[1010] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, p. 10.
[1011] SKITKA, L. J. *et al.*, "Automation bias and errors: are crews better than individuals?", *The International Journal of Aviation Psychology*, vol. 10, No. 1, 2000, p. 86.
[1012] DE-ARTEAGA, M. FLOGLIATO, R. & CHOULDECHOVA, A., "A case for humans-in-the-loop…", *cit.*, 2020, pp. 1-12.

make mistakes [1013] and sometimes even consciously reinforce exiting situations of oppression.[1014]

Thus, this approach to regulating algorithms follows from the fact that even though algorithms generally make better and less biased decisions than human decision-makers, they can still produce undesirable outcomes in new ways that need to be specifically addressed.[1015] It is thus necessary to regulate algorithmic decision-making in order to prevent and deal with certain unwanted situations, such as erroneous or discriminatory outcomes, generated by the use of these automated tools.[1016] KAMINSKI labels this approach to justifying the regulation of algorithms as "instrumental": "We should regulate algorithms, this reasoning goes, to prevent the consequences of baked-in bias and discrimination, and prevent other kinds of error".[1017]

This is particularly relevant when algorithms are used by public administrations in the provision of public services addressed towards redistributing wealth and protecting vulnerable populations.[1018] The risks generated by flawed algorithmic outcomes in these cases are even greater since the harms caused may mean worsening the conditions of individuals and groups that have been traditionally oppressed such as the poor and racial/ethnic minorities. Moreover, these individuals generally have fewer resources than the wealthy to access and activate redress mechanisms or even to detect the existence of errors. It is therefore essential to develop regulatory instruments or adapt existing ones in order to target automated system errors, in particular to prevent, control and deal with instances in which risks are generated for vulnerable or traditionally oppressed sectors of the population.

Since algorithmic decision-making generates problems that already existed, such as the perpetuation of structures of oppression, but in new ways, it is necessary to develop a regulatory framework that addresses and deals with the different innovative forms of producing undesirable outcomes that result from the use of automated decision-making tools. While traditional existing regulatory instruments are designed deal with the problems that may be contained and result from human decision-making, the flaws and problems that can

[1013] O' NEIL, *Weapons of Math Destruction*, *cit.*, 2016.
[1014] EUBANKS, V., *Automating Inequality…, cit.*, 2017, p. 38.
[1015] CHOULDENOVA, A., "Fair prediction with disparate impact…", *cit.*, 2016, p. 5
[1016] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1540.
[1017] *Ibidem.*
[1018] SCHARTUM, W., "Law and algorithms in the public domain", *Etikk i praksis, Nordic Journal of Applied Ethics*, No. 1, 2016, p. 16; EUBANKS, V., *Automating Inequality…, cit.*, 2017.

arise from automated decision-making are not addressed by equally comprehensive regulatory and institutional mechanisms.[1019]

1.2. OPACITY (LACK OF TRANSPARENCY)

The lack of transparency of automated systems does not exclusively extend to understanding how the processes carried out by these technologies work and what the logic underlying specific decisions is, but also includes the role that algorithms play in society[1020] and the specific purposes for which they are used. As information societies become more mature,[1021] their dependency on algorithms and other related technologies increases and their penetration is so high that both their role and presence is mostly either ignored or gone by unnoticed.[1022] Additionally, individuals are sometimes not even aware of the fact that they are being subjected to automated or semi-automated processing because the use of these systems is kept secret.[1023]

Machine learning algorithms are constructed by layers upon layers of programming and are therefore impossible to understand at user-level knowledge and even, in many cases, for experienced engineers.[1024] Moreover, once the algorithms are deployed they are constantly learning and evolving and it is therefore almost impossible to establish one single explanation for the way in which an automated system works that is kept true and accurate as it evolves after it has been released.[1025]

System opacity can operate in two ways: as general opacity, which takes place when the way in which algorithms work is kept secret, thereby preventing general system control and accountability and specific opacity, which refers to the way in which the lack of transparency of automated decision-making affects individuals and prevents them from effectively defending their rights and challenging algorithmic decisions.

[1019] YEUNG, K., "Why worry about decision-making by machine?", *cit.*, 2019 p. 23.
[1020] CATH, C. *et al.*, "Artificial intelligence and the 'good society'…", *cit.*, 2018, p. 507.
[1021] FLORIDI, L., "Mature information societies – a matter of expectations", *Philosophy and Technology*, vol. 29, No. 1, 2016a, pp. 1-4.
[1022] CATH, C. *et al.*, "Artificial intelligence and the 'good society'…", *cit.*, 2018, p. 507.
[1023] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, pp. 1091-1092.
[1024] MONASTERIO ASTOBIZA, A., "Ética algorítmica…", *cit.*, 2017, p. 188.
[1025] BURRELL, J., "How the machine 'thinks': understanding opacity in machine learning algorithms", *Big Data & Society*, vol. 3, No. 1, 2016, p. 5.

The way in which automated systems work is generally kept secret due to a wide variety of reasons that can range from trade secret protection to the prevention of gaming.[1026] This form of opacity hampers effective system oversight and control.[1027] If the way in which an automated system operates is not at least made available to oversight bodies or a mechanism through which it can be controlled is not provided, it is not possible to ensure that the system complies with Rule of Law values: accountability is not possible without a certain degree of transparency.

Specific system opacity takes place when an algorithm is used in a particular decision-making process. In said instances, system complexity leads individuals to lack knowledge regarding how the final decision that affects them was made, that is, how the system operates in specific cases.[1028] The chances that individuals have in challenging automated decisions can be heavily undermined if individuals who are subjected to automated processing are not provided with comprehensible explanations of the process and rationale behind the decision that affects them.

1.3. JUSTIFICATION

Even if and explanation is provided, the explanation provided with regard to the logic underlying the algorithmic decision-making system must be considered satisfactory.[1029] Providing an explanation does not necessarily mean that a decision is sufficiently justified.[1030] For example, if a student who has failed an exam asks the professor for an explanation for the decision made and the professor responds by saying that she has chosen to fail the student due to her resemblance with someone that failed that same exam the previous year, the explanation provided does clearly not justify the decision made. Hence, for a justification to be considered sufficient, it must not just be understandable, but the factors that lead up to it must be relevant,[1031] as BAROCAS and SELBST put it, they must be intuitive,[1032] that is, based on good reasons. Justifications are essential to provide decisions with

[1026] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, pp. 1092-1093.
[1027] MITTELSTADT, B. D. *et al.*, "The ethics of algorithms…", *cit.*, 2016, pp. 6, 10.
[1028] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, p. 1092.
[1029] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1545-1546.
[1030] BRENNAN-MARQUEZ, K., "'Plausible cause': explanatory standards in the age of powerful machines", *Vanderbilt Law Review*, vol. 17, No. 4, 2017, p. 1288.
[1031] OSWALD, M., "Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power", *Philosophical Transactions of the Royal Society of London*, *Series A: Mathematical and Physical Sciences*, vol. 376, No. 2128, *cit.*, 2018, p. 3.
[1032] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, p. 1091.

legitimacy, which is necessary for social systems to work. Justifications are also necessary when establishing due process rights systems and for the respect of human dignity; a decision that affects a human being is justified out of respect for that person, who is treated as a human and not as an object that can be used.

Humans have the capacity of introducing certain values and contextual knowledge to a particular decision and thus the ability of producing explanations that offer sufficient justification for the decision that has been made.[1033] Conversely, algorithms generally lack the ability to introduce said contextualised knowledge and, if certain specific commands are not fed to the algorithm during its development phases, it is highly likely that they will reach conclusions based on unexpected correlations that, although may actually be statistically accurate, seem to be random and unreasonable,[1034] and thereby not properly justified.

Hence, in order to provide legitimacy to the decision-making system and to respect individuals' dignity and rights, it is essential to set up a regulatory framework that establishes a series of requirements regarding the underlying logic and justification of the individualised decision made by the algorithm.[1035] This does not necessarily mean providing full justifications for all automated decisions but weighing in which cases traditional forms of justification are still necessary and in which cases it is more convenient to forgo full explainability and intuitiveness for the sake of efficiency. In the second case, it will be necessary to develop effective oversight systems for opaque algorithmic decision-making processes.

1.4. RISKS TO DIGNITY: INDIVIDUALITY, AUTONOMY AND PRIVACY

Concerns regarding the use of algorithms based on notions of dignity hold that treating individuals exclusively through automated tools dehumanises them.[1036] By treating someone

[1033] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1546.
[1034] FELTEN, E., "What does it mean to ask for an 'explainable' algorithm?", *Freedom to Tinker*, 31st May 2017. Available on 9th July 2019 at: https://freedom-to-tinker.com/: "…imagine that an algorithm for making credit decisions considers the color of a person's socks, and this is supported by unimpeachable scientific studies showing that sock color correlates with defaulting on credit, even when controlling for other factors. So the decision to factor in sock color may be justified on a rational basis, but many would find it unreasonable, even if it is not discriminatory in any way."
[1035] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1549.
[1036] JONES, M. L., "The right to a human in the loop: Political constructions of computer automation and personhood", *Social Studies of Science,* vol. 47, No. 2, 2017, pp. 231-232.

as a thing, they stop being an end to become a means towards an end for the user, thereby being deposed of their personhood, identity and dignity.[1037]

Dignitary arguments are especially valid when applied to the growing use of automated systems in welfare programmes, which risk treating members of vulnerable groups as measurable objects. There is a growing risk that algorithms used to make better decisions in welfare programmes shift their focus from providing the best possible aid to vulnerable populations to controlling that there is no fraud in the actions of welfare beneficiaries.[1038] Hence, the main objective of welfare programmes would cease to be providing vulnerable populations with a minimum safety net based on notions on solidarity (not charity), thereby recognising and respecting their dignity and social responsibility in those life contexts that can lead to economic needs. This shift in aims would contribute to reinforcing negative stereotypes of the poor, undermining the dignitary dimension of policies that should be built in order to attain substantive equality objectives.[1039]

There are also other concerns that derive from the need to protect human dignity. These include risks to autonomy and individuality and harms to fundamental rights, in particular, privacy.

Automated systems tend to make decisions by grouping individuals and making generalisations. The algorithm infers certain group membership from characteristics it detects in the individual and, once, said group membership has been inferred, it also ascribes previously unknown traits to the individual that are typical of members of the group she belongs to or is associated with.[1040] Since the individual's particular traits are extracted from the correlations developed between her data and the data of other similar individuals, she is not being treated as an individual but as part of and in regard to the group.[1041]

This argument closely relates to the criticism that applies to legal rules that are built from the perspective of the prototypical liberal individual.[1042] When norms are not constructed and applied in a contextualised manner it is highly likely that they will lead to damaging

[1037] KANT, I., *Grounding for the Metaphysics of Morals*, Indiannapolis, Hackett Publishing Company, 3rd ed, 1993 (first pusblished in 1785), p. 36.
[1038] RANCHORDAS, S. & SCHUURMANS, Y., "Outsourcing the welfare state…", *cit.*, 2020, pp. 14-19.
[1039] FREDMAN, S., *Discrimination Law*, *cit.*, 2011, pp. 28-30.
[1040] ZARSKY, T., "Transparent predictions", *University of Illinois Law Review*, vol. 2013, No. 4, 2013, pp. 1560-1561.
[1041] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1542.
[1042] BARRANCO, M. C., *Diversidad de Situaciones y Universalidad de los Derechos*, *cit.*, 2011, pp. 14-15.

outcomes for members of disadvantaged groups. This comparison might seem counterintuitive to the extent that the arguments against the uniform application of rules mainly defend the need to consider the particular traits of individuals as members of a disadvantaged group and thus argue that members of disadvantaged groups should be considered in relation to the group. However, far from eliminating individuality, both arguments pointed out aim to correct negative profiling resulting from a failure to consider an individual's particular characteristics either when these differ from the negative attributes associated to a certain population group[1043] or when precisely the individual's membership to a particular group means that a set of specific traits (such as cultural traditions) that will benefit the individual with regard to a decision must be considered.[1044]

Additionally, the computational capacity automated tools have, which provides amongst other things, the possibility of creating profiles that will self-develop, generates important risks not only for individuality but also for autonomy. Once the individual loses the capacity of deciding the way in which her digital personality is presented and that power passes on to either external entities or, in the context of the research here developed, to machine learning algorithms, not only does she lose her individuality, but also her autonomy.[1045] In order to preserve individuality and autonomy and prevent possible situations of objectification that derive from the creation of digital profiles it is thus necessary to establish mechanisms that provide individuals with the possibility of challenging and introducing changes in their profiles.[1046]

Autonomy concerns also arise with regard to possible instances of digital manipulation. When organisations make decisions based on automated tools they can manipulate individuals into adopting attitudes or forming their identity in a certain way that restricts their autonomy.[1047] A clear example of this is targeted advertising, through which firms decide who will view the advertisements they post, thereby excluding certain groups of people from

[1043] BOHREN, J. A. *et al.*, "Inaccurate statistical discrimination", *cit.*, 2019.
[1044] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1542-1543: "Take, for example, the USDA algorithm that decided that Somali groceries in Seattle would no longer be permitted to accept food stamps because purportedly suspicious customer behavior indicated cash-for-stamp fraud. The algorithm responded to "suspicious transactions" such as even dollar amounts and large purchases made in short time spans. But the grocers had credible, more individualized, explanations for these purportedly "unusual" practices: the shopping patterns of East African immigrants, which include shopping in groups, and shopping for meat by whole dollar amounts at Halal butchers".
[1045] BYGRAVE, L. A., "Minding the machine: Article 15 of the EC data protection directive and automated profiling", *Computer Law & Security Review*, vol. 17, No. 1, 2001, p. 18.
[1046] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1543-1544.
[1047] *Idem*, pp. 1544-1545.

easily accessing their content.[1048] Taking this kind of action undeniably restricts individuals' autonomy by limiting their possibility of freely deciding whether they want to purchase a certain product or not or if they want to apply for a certain position.[1049] Additionally, the level of manipulation that is employed in predatory advertising poses serious risks, not only to individual autonomy in general but particularly regarding the free will of members of groups at risk of social exclusion.[1050] Manipulation through advertising is not new. However, recent and ongoing technological developments offer organisations a series of possibilities in the creation of choice architectures that allow for highly personalised environments that can influence and alter individuals' behaviour at unprecedented levels through "hypernudging".[1051]

With regard to the harms that algorithms can cause to fundamental rights, special attention has been paid to privacy, as a right that directly stems from the protection of human dignity. As it will be discussed later on, privacy harms are the most obvious harms caused to individuals through the use of algorithms, which basically are data processing technologies. The invasion of an individual's privacy by collecting and processing their data without them fully understanding and consenting to said actions and the purposes for which collection and processing are carried out, is constitutive of a direct harm to the person's dignity. Personal data is, after all, information on how an individual's identity is built. Attacks to the sphere of informational self-determination, thus cause harms to the very core an essence of an individual's dignity.[1052]

1.5. PARTICIPATION AND DUE PROCESS

In general, individuals are not given many opportunities of participating in the automated decision-making process and contesting decisions. The existence of rights and mechanisms that guarantee individuals the possibility of intervening in the decision-making process becomes especially relevant when said process cannot be easily explained.[1053] However, the shift from traditional to automated decision-making is not always accompanied by a set of

[1048] DATTA, A., TSCHANTZ, M. C. & DATTA, A., "Automated experiments on ad privacy settings…", *cit.*, 2015, p. 102.
[1049] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1544-1545.
[1050] O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, p. 71.
[1051] YEUNG, K., "'Hypernudge': Big data as a mode of regulation by design", *Information, Communication & Society*, vol. 20, No. 1, 2017, pp .118-136.
[1052] FLORIDI, L., "On human dignity as a foundation for the right to privacy", *Philosphy & Technology*, vol. 29, 2016b, pp. 307-312.
[1053] YEUNG, K., "Why worry about decision-making by machine?", *cit.*, 2019, p. 25.

procedural rights that provide individuals with the possibility of challenging the algorithmic decisions that affect them. Some of the cases that have been studied by the scholarship show a general reluctance not only to offer explanations of the decisions made by algorithms but also to open up mechanisms to contest decisions.[1054]

In fact, precisely due to the generalised belief that algorithms are better decision-makers, individuals are in many cases offered less possibilities of defending their position and challenging decisions that they would if a human being were making the decision.[1055] Furthermore, even when individuals manage to, for instance, challenge erroneous profiles, the processes that must be undergone in order to prove that the information or predictions made by the algorithm are mistaken are very lengthy and costly.[1056] Additionally, by the time the algorithm is proven to have made a mistake, the implications that it has had for the affected individuals may be impossible to reverse.[1057] For example, when programmes erroneously identify a passenger or even a crew member as a terrorist they are not permitted to fly and thus by the time their name is cleared, after being generally subjected to questioning for hours, they will have probably missed their flight.[1058]

Due process rights provide individuals with the possibility of defending their position and challenging decisions before a neutral arbiter, thereby ensuring that there is a chance of controlling the legality of the decision. The mere existence of rights that individuals can exercise in order to challenge decisions and activate a process that ensures that decisions respect the rules and regulations applicable to them increases the perceived fairness and the legitimacy of decision-making systems.[1059] Although legitimacy is not as important in private decision-making as it is in public decision-making, it is also generally relevant to ensure that decisions (and the whole decision-making system in which they are inserted) are taken seriously and respected. These rights should not operate exclusively as *ex post* mechanisms that can be activated once the decision has been made, individuals should also be provided with the possibility of participating throughout the whole decision-making process.

[1054] O'NEIL, C., *Weapons of Math Destruction…*, *cit.*, 2017, pp. 3-11.
[1055] KAMINSKI, M. E., "Binary governance…", *cit*., 2019b, pp. 1538-1539.
[1056] CITRON, D. K., "Technological due process", *cit.*, 2008, p. 1273: "…the FPLS identified the wrong man in a case involving a $206,000 child-support debt. It took the accused man and his attorney over two months to convince the California district attorney's office that the system had made a mistake."
[1057] This particular implication of algorithmic decision-making can also be true for human decision-making but it is, nonetheless, still important to point it out it in order to highlight the need for an effective system of "technological due process rights".
[1058] CITRON, D. K., "Technological due process", *cit.*, 2008, p. 1274-1275.
[1059] KAMINSKI, M. E., "Binary governance…", *cit*., 2019b, pp. 1547-1548.

As with transparency, participation operates on both a general and a specific or individual level. The individual level refers to particular algorithmic decision-making processes. The general level refers to broad participation in the design of a society in which an increasing number of processes are automated. Moreover, general channels for participation in the design and creation of algorithms and the ways in which they are used should also be opened.[1060] The expansion of automated systems is generating a complete paradigm shift in the way that many forms of social, political and economic relations are shaped and carried out. Additionally, these systems significantly influence power structures. It is therefore only logical that individuals and groups of citizens are offered the possibility of participating and deciding how they want their technological future to be structured.[1061]

Finally, failure to guarantee technological due process and participation rights is also closely related to dignitary concerns regarding automated processing and decision-making. Due process and participation rights, such as the possibility to be heard during the process and contest decisions, provide the necessary tools in order to treat individuals as active subjects with moral agency.[1062] If individuals are denied such rights, respect for human dignity will be undermined. In order to restore the humanity that individuals are deposed of when being subjected to algorithmic decision-making it is necessary to establish a set of regulatory mechanisms that guarantee that individuals' dignity is considered in the decision-making process.[1063]

1.6. TRACEABILITY

Closely related to due process rights is the need to attribute responsibility for undesirable, unfair, erroneous or discriminatory outcomes. Machine learning algorithms are constantly learning from their environment and evolving. The individuals who create these systems can sometimes claim to only be partially responsible for the decisions and conclusions drawn by the algorithm once it has been deployed, for they could not have foreseen all possible outcomes. For this reason it is highly difficult to directly determine who is responsible and, if

[1060] YEUNG, K., "Why worry about decision-making by machine?", cit., 2019, p. 26.
[1061] *Ibidem.*
[1062] YEUNG, K., "Why worry about decision-making by machine?", *cit.*, 2019, p. 26.
[1063] JONES, M. L., "The right to a human in the loop…", *cit.*, 2017, pp. 231-232.

necessary, liable, for the decision made by the algorithm. This is particularly problematic when dealing with controversial applications such as Lethal Autonomous Weapons.[1064]

The high degree of complexity of these systems means that, even if a human is assigned the task of supervising the automated systems used and is given power to override decisions, she might not be able to fully grasp or comprehend the whole extent of the machine's process, nor control it.[1065] When automated systems reach such a degree of complexity it might sometimes be difficult to identify who should be blamed for erroneous, unfair or discriminatory outcomes.

This "responsibility gap" [1066] has appeared in some cases of algorithmic discrimination as it has been impossible to detect were the discriminatory result was originally generated.[1067] As it is to be expected, firms such as Google have tried (and generally succeeded) to declare their lack of responsibility when search engine results help to perpetuate structures of discrimination. [1068] Hence, only if regulatory instruments establish the way in which responsibility should be attributed in cases in which algorithms produce harms, it will be possible to make sure the design and use of algorithms in processes that directly or indirectly affect individuals is properly controlled.[1069]

1.7. THE LEGITIMACY AND LEGALITY OF PUBLIC AUTOMATED DECISION-MAKING

Legitimacy is essential for the survival of political systems. The authority held by political leaders and the respect for political and normative systems is severely undercut when populations no longer consider them legitimate.[1070] In democracies, a very important part of the system's legitimacy is founded on the establishment of procedural rules. Procedural rules grant individuals the necessary safeguards and knowledge to defend their rights against any

[1064] CATH, C. *et al.*, "Artificial intelligence and the 'good society'…", *cit.*, 2018, p. 511; HUMAN RIGHTS WATCH, "Mind the gap: the lack of accountability of killer robots", 2015.
[1065] MITTELSTADT, B. D. *et al.*, "The ethics of algorithms…", *cit.*, 2016, p. 6.; YEUNG, K., "Why worry about decision-making by machine?", *cit.*, 2019, p. 25.
[1066] HUMAN RIGHTS WATCH, "Mind the gap…", *cit.*, 2015.
[1067] DATTA, A., TSCHANTZ, M. C. & DATTA, A., "Automated experiments on ad privacy settings…", *cit.*, 2015, pp. 92-112; SWEENEY, L., "Discrimination in online ad delivery", *cit.*, 2013, pp. 44-54.
[1068] NOBLE, S. U., *Algorithms of Oppression…*, *cit.*, 2018, p. 42: "Google's assertion that its search results, though problematic, were computer generated (and thus not the company's fault) was apparently a good-enough answer for the Anti-Defamation League (ADL)…".
[1069] The GDPR does, to a certain extent, establish a system of liability for processors and controllers (see, for instance, article 82).
[1070] TYLER, T. R., *Why People Obey the law*, New Haven and London, Yale University Press, 1990, p. 5.

decision made by public authorities that affects them.[1071] Even if an administrative or judicial process results in a negative outcome for an individual, if procedural rules are, overall, perceived to be fair, decisions will generally be accepted once they are final.[1072] That is, if a system of Rule of Law is correctly established and guarantees that all decisions made respect the regulatory instruments applicable in each specific case, said decisions and those who make them will be considered legitimate.[1073] Legitimacy is thus a by-product of the system's perceived fairness, which is largely constructed through the existence of procedural rights, such as the right to a fair trial and effective remedy.

Adapting existing decision-making systems to the new realities brought by automated processing tools in order to provide them with legitimacy is closely linked to the need to adapt the principles and procedures of public administrations to the new reality of algorithmic decision-making. Establishing mechanisms that ensure the system's legitimacy and the effectiveness of procedural rights is especially important when it comes to the use of algorithms by the public sector. After all, public bodies are especially bound by fundamental rights and democratic principles. In addition to the general problems or concerns that arise with regard to the use of data processing technologies that have been examined thus far, there are a series of specific concerns that apply to the public use of automated systems which result from the private exercise of inherently public tasks that sometimes comes with automation and the increased demands for transparency and justification that in theory apply to public decisions and that, in many cases, are not met when automated systems are used.

1.7.1. The private exercise of inherently public tasks

Although the externalisation of public tasks is not new, externalisation through the use of private algorithms brings about specific concerns, especially in sectors of public activity that involve the use of coercive powers.

An excessive reliance on algorithmic systems can lead to very significant shifts in the way in which administrative discretionary powers are applied. The impossibility of foreseeing every possible situation that will come up within a regulated area means that policy implementation and application of rules to specific cases sometimes depends on the discretionary powers that

[1071] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, p. 1119.
[1072] TYLER, T. R., *Why people obey the law*, *cit.*, 1990, p. 7.
[1073] YEUNG, K., "Algorithmic regulation: a critical interrogation", *Regulation & Governance*, vol. 12, No. 4, 2018, p. 517.

are granted to administrative bodies, units and public servants.[1074] Attributing discretionary powers to public bodies means that the application and control of legality is entrusted to certain individuals and organisations that are vested with public authority. If algorithms are used in ways and for purposes that effectively entail overtaking the decision power of humans vested with public authority, the actual decision-makers will not be the public servants that are only supposed to be using automated systems as support tools, thereby modifying and reducing the actual decision-making power of public authorities.[1075]

A closely related issue regarding the use of algorithms by the public sector results from the fact that these tools are generally bought from or developed alongside private firms.[1076] This leads to a whole new framework in the interaction between private and public spheres of activity and draws questions on the extent to which these private firms are exercising administrative discretionary powers and intervening in policy-making and agenda-setting activities which should be carried out by the people's elected representatives and public servants vested with public authority.[1077] Traditional regulatory instruments are general and ambiguous in order to allow for their adaptation and application to particular cases. Conversely, algorithms need to be programmed in a very specific manner. Although the set of instructions that is the algorithm will vary once it is applied to specific cases, the initial level of granularity that algorithmic programming requires, entails a degree of precision in the instructions provided that is not used when creating legal code. Hence, the individuals and firms creating algorithms will be interpreting and applying legal instruments in the very specific manner that would theoretically correspond to public representatives or servants.[1078] Additionally, in some cases, AI tool developers allow the use of their technologies by the public sector in exchange for data,[1079] which brings about concerns regarding the protection of personal data.

[1074] OSWALD, M., "Algorithm-assisted decision-making in the public sector…", *cit.*, 2018, p. 14.
[1075] *Idem*, pp. 14-15.
[1076] BRAUNEIS, R. & GOODMAN, E. P., "Algorithmic transparency for the smart city", *Yale Journal of Law & Technology*, vol. 20, 2018, p. 113.
[1077] VEALE, M. & BRASS, I., "Administration by algorithm", in YEUNG, K. & LODGE, M., (eds.), *Algorithmic Regulation*, Oxford, Oxford University Press, 2019, p. 133.
[1078] *Ibidem.*
[1079] SHEAD, S., "Google DeepMind is giving the NHS free access to its patient monitoring app", *Business Insider*, 24th June 2017. Available on 5th April 2019 at: https://www.businessinsider.de/

1.7.2. Transparency and justification of public decisions

Establishing mechanisms and tools aimed towards eliminating algorithmic opacity becomes particularly relevant when addressing public automated decision-making seeing as the principle of transparency plays a vital role in democratic Administrative law systems.[1080] There are several ways in which transparency advances the democratisation of societies. Firstly, transparency empowers citizens as it provides them with information on the actions and procedures carried out by public powers enabling them to participate in processes which may interest them. Secondly, and closely related to the previous point, for public institutions to be accountable, it is necessary not only to establish procedural rules that allow individuals to defend their position and rights, but to provide said individuals with the necessary knowledge regarding the way in which a decision that affects them and which they may want to challenge has been made.[1081] Thirdly, transparency is also necessary in order to provide individuals and organisations with the necessary knowledge to contest and start public debates regarding any actions or decisions made by public powers and, thus, to exercise effective oversight over said actions. Finally, transparency provides legitimacy to public institutions. If citizens know what actions and decisions that, directly or indirectly affect them, public powers are making and how to challenge them and participate in public processes, they will increase their trust in the system.

As it was already indicated in the section dedicated to algorithmic opacity, in order to implement an effective system for the control and oversight of algorithmic tools, certain levels of transparency are necessary. Said transparency must not solely refer to making the source code public but, in certain cases, also to providing explanations that are understandable.[1082] The need for explanations is especially important in the public sector given the fact that the decisions carried out by public bodies must be justified, that is, the elements considered must be relevant for the decision made.[1083] Consequently, it will be necessary either to ensure that public automated decision-making is capable of producing justifications that are acceptable within the current regulatory framework or to move away

[1080] MESTRE DELGADO, J. F., "Una reflexión sobre la regulación constitucional del Derecho administrativo", *Corts: Anuario de Derecho Parlamentario,* No. Extra 31, 2018, p. 384.
[1081] COGLIANESE, C. & LEHR, D., "Regulating by robot…", *cit.*, 2017, p. 1205; ANNANY, M. & CRAWFORD, K., "Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability", *New Media & Society*, vol. 20, No. 3, 2018, p. 974.
[1082] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, p. 1091.
[1083] VEALE, M. & BRASS, I., "Administration by algorithm", *cit.*, 2019, p. 131.

from what as of now is considered a sufficient justification towards accepting nonintuitive explanations and ensuring accountability in other ways.[1084]

The specific problems that the use of machine learning algorithms by the public sector brings by[1085] must however not become a deterrent for the use by public bodies of the technologies here analysed.[1086] If public institutions want, as it is their duty, to keep up with the private sector in order to effectively regulate it when necessary, they must also employ machine learning and other related technologies that help them to make better decisions.[1087]

2. TRADE-OFFS IN THE REGULATION OF ALGORITHMS

There are many regulatory solutions that have been put forward in order to solve the issues that have been pointed out in the previous section. As with any regulatory proposal, when adopting these specific solutions to deal with problems such as algorithmic opacity or lack of accountability, there are several trade-offs that regulators will encounter both in the public and private sectors.

Analysing the trade-off between different sets of interests becomes especially relevant with regard to the use of algorithms by public institutions. In these cases, we face a conflict between prioritising the interests of those being subjected to automated decision-making and general public interests.

One of the most relevant trade-offs that takes place is the one that arises from the tensions between transparency and public security which includes, amongst other elements, state secrets[1088] and non-disclosure to avoid gaming in law enforcement.[1089] The latter is clearly exemplified when dealing with transparency in the criminal justice system. The extent to which the fundamental rights of individuals can be affected by the use of algorithms in law

[1084] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, p. 1138.
[1085] COGLIANESE, C. & LEHR, D., "Regulating by robot…", *cit.*, 2017, pp. 1151-1152; CATH, C. *et al.*, "Artificial intelligence and the 'good society'…", *cit.*, 2018, pp. 511-512.
[1086] COGLIANESE, C. & LEHR, D., "Regulating by robot…", *cit.*, 2017, p. 1153.
[1087] COGLIANESE, C., "Optimizing regulation for an optimizing economy", *University of Pennsylvania Journal of Law & Public Affairs*, vol. 4, No. 1, 2018, p. 11.
[1088] BRKAN, M., "Do algorithms rule the world? Algorithmic decision-making in the framework of the GDPR and beyond", 2017, pp. 23-24. Available on 9th May 2019 at: https://ssrn.com/ (the more recent version of this paper, published in the *International Journal of Law and Information Technology*, vol. 27, No. 2, and which will be cited later on, does not address the constraints to algorithmic transparency cited here).
[1089] BAYAMLIOĞLU, E., "Transparency of automated decisions in the GDPR: an attempt for systemisation", 2018, p. 18. Available on 9th May 2019 at: https://ssrn.com/; ZARSKY, T., "Transparent predictions", *cit.*, 2013, pp. 1553-1554.

enforcement means that the automated processes they are subjected to should be made public and available to them. However, the tensions that arise between conflicting interests in these cases are especially intense given the fact that the elements that must be weighed are, on the one hand, the fundamental rights of alleged offenders and the democratic quality of the judicial system and, on the other hand, public security.

Disclosing the underlying logic of algorithms used in law enforcement and the criminal justice system could provide offenders with the necessary tools to game the system and avoid being caught.[1090] Consequently, especially when regulating the use of automated systems in law enforcement it is necessary to carefully examine how much transparency can be granted in order to guarantee individual rights of those under investigation or prosecution while also preventing offering enough information to enable gaming.

Another trade-off that has to be considered is the one that arises with regard to algorithmic understandability. Source code transparency is ineffective if its complexity renders it meaningless. For a system of algorithmic accountability based on transparency to be effective, automated systems must be understandable.[1091] The problem is that making algorithms understandable can sometimes require simplifying them and reducing their accuracy.[1092]

The way in which the conflict between transparency and other interests plays out in the private sector is very different for there is no general transparency principle that underpins private firms' activities. In addition, imposing algorithmic transparency obligations on private entities is one of the main reasons why certain sectors of the scholarship argue that regulation stifles innovation.[1093] IP law, including trade secrets, aims to foster creativity and productivity in order to maximise social welfare by protecting the innovations produced by individuals or firms.[1094] By ensuring individuals and businesses that their creations will be correctly protected and that they will reap most of the benefits that result from them, they are incentivised to invest more resources in order to continue with innovative processes.[1095]

[1090] BAYAMLIOĞLU, E., "Transparency of automated decisions in the GDPR…", *cit.*, 2018, p. 18.
[1091] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1548.
[1092] OSWALD, M., "Algorithm-assisted decision-making in the public sector…", *cit.*, 2018, p. 4.
[1093] YEUNG, K. & LODGE, M., "Algorithmic regulation…", *cit.*, 2019, p. 7.
[1094] WORLD INTELLECTUAL PROPERTY ORGANISATION, "What is intellectual property?", 2004, p. 3. Available on 22nd May 2019 at: https://www.wipo.int/
[1095] STOPFAKES.GOV (INTELLECTUAL PROPERTY RIGHTS INFORMATION & ASSISTANCE), "Why is intellectual property important?", 7th July 2016. Available on 22nd May 2019 at: https://www.stopfakes.gov/: "Intellectual property protection is critical to fostering innovation. Without protection of ideas, businesses and individuals

Consequently, if private firms are forced to reveal proprietary elements of the systems they have developed, incentives to innovate and advance technological development will be reduced.

Other forms of intervention or obligations imposed to the private sector, such as imposing systems of authorisations or prior certifications, can also be argued to disincentivise innovation. In this regard, placing mechanisms such as authorisations systems that must be complied with before a product or service enters the market poses obvious burdens on goods and services providers and reduces their economic profits.

Conflicts also arise with regard to the extent to which it is possible to regulate and limit firms' freedom of enterprise. For example, several scholars have pointed out that the reason why algorithms yield discriminatory results and reproduce traditional narratives of oppression is that most of the programmers who work in the tech sector are white males.[1096] As it was already indicated in part I, possible regulatory choice to help solve this problem is to force firms to hire a more diverse workforce, a choice that may be argued to clash with firms' right to self-organisation. Regulators will therefore have to choose between limiting firms' right to self-organisation in order to fast-track diversity in the tech sector or to trust that the measures that said companies adopt will eventually lead to a more egalitarian workforce and, in turn, less discriminatory algorithms.

Nonetheless, some of the measures that should be carried out entail very low levels or no intervention in private technological firms. This is the case of the measures that should be taken in the educational system in order to address the elements that make girls and women not choose or drop out of STEM degrees. For instance, initiatives such as "black girls code", which is a programme that offers courses specifically addressed to teaching black girls how to code[1097] help to promote female and minority engagement in STEM subjects and their pursuit of careers in science and technology.[1098] Additionally, it is also essential for ethics courses to be introduced in engineering degrees in order to teach students in these fields the

---

would not reap the full benefits of their inventions and would focus less on research and development. Similarly, artists would not be fully compensated for their creations and cultural vitality would suffer as a result."

[1096] NOBLE, S. U., *Algorithms of Oppression…*, *cit.*, 2018, p. 80.

[1097] BLACK GIRLS CODE, "What we do", 2018. Available on 13th June 2019 at: http://www.blackgirlscode.com/

[1098] NOBLE, S. U., *Algorithms of Oppression…*, *cit.*, 2018, p. 64.

ways in which algorithms and programming can lead to discriminatory outcomes, hence making them aware of the risks that can arise when developing software programmes.[1099]

Regulators will also face a series of trade-offs when deciding what requirements and restrictions should be set for algorithms and the standards to which different types of automated systems should be held to. For example, using certain types of algorithmic systems or using them for certain purposes might be deemed too risky and thus be altogether prohibited. This is the context in which we can, for instance, find article 22 of the GDPR. Article 22 bans solely automated decision-making when it has legal or significantly similar effects on individuals. Drawing from this idea, rules may also be passed in order to prohibit contents which perpetuates negative stereotypes from being associated and shown when certain words are introduced in search engines.[1100]

Very complex choices are also derived from the increased efficiency that algorithms provide processes with. An example that was already pointed out in part I, is the trade-off between equality and efficiency that will take place when deciding whether to allow algorithmic decisions which, although accurate, result from structural discrimination and reinforce pre-existing situations of oppression.[1101] Privacy rights also enter into conflict with algorithmic efficiency and accuracy due to the fact that prohibiting the use of certain types of data will mean that automated systems will not have all the necessary information to draw accurate profiles.[1102]

However, a more pressing issue than the conflict between privacy and accuracy is that the very framework of protection set by informational privacy may not be useful any longer.[1103] Firstly, prohibiting the use of certain types of data does not only lead to a loss in general efficiency, but also increases the possibilities that wrong inferences regarding individuals

[1099] NOBLE, S. U., *Algorithms of Oppression…*, *cit*., 2018, pp. 69-70; O'NEIL, *Weapons of Math Destruction…*, *cit.*, 2017, pp. 145-146: "...people ask me how to teach ethics to a class of data scientists. I usually begin with a discussion of how to build an e-score model and ask them whether it makes sense to use 'race' as an input in the model. They inevitably respond that such a question would be unfair and probably illegal. The next question is whether to use 'zip code'. This seems fair enough, at first. But it doesn't take long for the students to see that they are codifying past injustices into the model"; WILK, A. "Teaching AI, ethics, law and policy", 24th May 2019, pp. 1-7. Available on 15th June 2019 at: https://arxiv.org/

[1100] NOBLE, S. U., *Algorithms of Oppression…*, *cit*., 2018.

[1101] CHOULDENOVA, A., "Fair prediction with disparate impact...", *cit.*, 2016; KLEINBERG, J., MULLAINATHAN, S., & RAGHAVAN, M., "Inherent trade-offs in the fair determination of risk scores", *cit.*, 2016, p. 5; GROSS, S. R., POSSLEY, M. & STEPHENS, K., "Race and wrongful convictions in the United States", *cit.*, 7th May 2017; BERTRAND, M., MULLAINATHAN, S. & ABRAMS, D., "Discrimination in the judicial system", *cit.*, 2001.

[1102] CUKIER, K., and MAYER-SCHOENBERGER, V., "The rise of big data…", *cit.*, 2013, p. 33.

[1103] SCHERMER, B. W., "The limits of privacy in automated profiling and data mining", *cit.*, 2011, p. 49.

may be made.[1104] Secondly, even if certain information, such as data concerning an individual's health,[1105] is hidden from the algorithm this will not necessarily prevent it from drawing inferences from other pieces of data from which the automated system will be able to create profiles containing, for instance, health information.[1106] Consequently, the trade-off between the right to privacy (data protection) and other rights, such as the right to equality, must be considered. This part is dedicated to analysing the information privacy framework as the regulatory tool that is currently available to deal with the issues and concerns discussed in a comprehensive manner.

3. THE PRIVACY FRAMEWORK AS A SOLUTION FOR THE HARMS CAUSED BY ALGORITHMS

Many of the real or potential harms caused by algorithms are the result of organisations collecting and processing the personal data of individuals. Consequently, instruments built from the perspective of the informational privacy protection framework are the most straightforward legal mechanisms to protect individuals from the harms caused by data processing technologies. In turn, the most obvious way in which to protect harms to privacy is through private law mechanisms, such as contract law, which is mainly built upon notions of consent.

However, many of the harms caused by the collection, processing, use and dissemination of personal data are public[1107] for they affect fundamental rights and basic democratic principles that are, in many cases, not internalised by the parties to data transactions. Moreover, as BEN-SHAHAR points out, not only does the "data market" suffer from negative externalities, but the lack of transparency that characterises data processing systems and individuals' general failure to perceive the real costs associated with sharing personal data in comparison with the obvious and short-term gains that are to be obtained when providing said data (for example, in order to access a website's content), lead to the existence of informational asymmetries and imperfect rationality.[1108] These market failures prevent private law mechanisms from being effective when dealing with the issues generated by the data services sector and, more specifically, by data processing technologies.

[1104] *Ibidem*.
[1105] Processing of personal data concerning health is prohibited by article 9 of the GDPR.
[1106] SCHERMER, B. W., "The limits of privacy in automated profiling and data mining", *cit.*, 2011, p. 49.
[1107] BEN-SHAHAR, O., "Data pollution", *cit.*, 2019, pp. 110-118.
[1108] *Idem*, pp. 121-124.

In this context, in which the protection of informational privacy from the risks generated by algorithms produces tensions between private and public law frameworks, the European Union has developed its data protection legal structure. The data protection framework, which is examined at large in this part, aims to provide a regulatory framework for many of the different concerns that arise from the current (and future) state of development of data processing technologies. Hence, while European data protection legal instruments are built from the perspective of privacy, the EU has, to a certain extent, acknowledged the public nature of some of the harms caused by the use of personal data.

The EU personal data protection regime thus places informational self-determination at its core by mostly relying on a framework that, in theory, provides individuals with authority over how and when to share their data. Nonetheless, this individual rights regime, which is heavily influenced by private law constructions, also establishes a series of mandates that controllers and processors must in all cases comply with when processing personal data. This means that some of the obligations comprised within the rights recognised to individuals are not just activated upon request by the data subject but must always be carried out by the physical or legal person processing or using the data. For example, articles 13 and 14 of the GDPR establish information disclosure obligations for controllers. Additionally, a series of public agencies have been set up in order to ensure public and private compliance with the right to data protection.

In addition, the EU personal data protection system also sets out a series of oversight tools that developed through combining soft law with some command-and-control regulatory tools in order to produce a system that aims to prevent and deal with the risks and harms caused by the processing of personal data. Finally, by establishing a network of data protection public authorities clearly conveys how the EU recognises the insufficiencies of exclusively tackling the hazards generated by data processing technologies through traditional private law instruments.

3.1. THE RIGHT TO DATA PROTECTION AS AN ANTI-CLASSIFICATION INSTRUMENT

Since this dissertation focuses on the rights to equality and non-discrimination, it is important to pay attention to the ways in which the rights to equality and non-discrimination interact and are related to the right to data protection and the privacy framework.

Both the GDPR and the Directive for data protection in law enforcement, which will also be briefly analysed in this section, make explicit references to the risk that data processing may lead to discriminatory outcomes.[1109] Moreover, while the GDPR only contains said references in its Recitals, the Directive does state, in article 11, that "profiling that results in discrimination against natural persons on the basis of special categories of personal data" is prohibited.

Considering decisions based on data processing are generally regulated by privacy regulatory instruments, in which data protection is an expression of informational privacy, the provisions in said regulatory instruments that are aimed towards preventing discrimination are mainly drawn from the perspective of anti-classification. The objective of regulatory instruments adopted in the EU is to approach the problems generated by automated decision-making and profiling through the adaptation of existing privacy regulations and the creation of new privacy instruments that specifically address the new issues generated by data processing technologies.[1110]

Not only do these instruments aim to protect the right to privacy as a valuable aspect of individuals' lives in its own standing, but these regulations also structure the right to privacy or, more specifically, to data protection, as an instrumental right through which to protect the right to equality and non-discrimination. However, as the following chapters convey, this approach to protecting equality and non-discrimination is insufficient and must be complemented with the anti-discrimination framework and, especially, with new regulatory instruments that specifically focus on preventing and dealing with the risks that automated systems generate with regard to the perpetuation of existing structures of oppression and disadvantage.

Since privacy instruments operate by prohibiting access, collection and processing of sensitive data (article 9 of the GDPR), this form of antidiscrimination regulation operates as a prevention mechanism and, consequently, does not solely focus on regulatory responses once the specific instances of discrimination have taken place.[1111] These regulations also provide victims of discriminatory practices with more tools to defend themselves seeing as they can

[1109] Recitals 71, 75 and 85 GPDR and recitals 23, 38, 51 and 61 and article 11 of the Directive for data protection in law enforcement.
[1110] NICHOLSON PRICE II, W. *et al.*, "Shadow health records meet new data privacy laws", *Science*, vol. 363, No. 6426, 2019, p. 448.
[1111] ROBERTS, J. L., "Protecting privacy to prevent discrimination", *cit.*, 2015, p. 2122.

base their claims on the fact that the decision was made using illegally obtained information.[1112]

Protecting informational privacy as a way in which to prevent algorithmic discrimination makes special sense due to the fact that the problems generated by machine learning algorithms result from their computational capability when processing information. Thus, protecting individuals' data and offering a catalogue of rights that provide them with the possibility of protecting their personal information, its accuracy and the decisions that result from its processing, is aligned with the concerns generated by existing information processing capabilities.

Choosing to focus on privacy over traditional anti-discrimination instruments in the prevention of algorithmic discrimination is also a logical option if we analyse Western regulatory traditions. Western legal culture mostly relies on the construction of individual legal categories, which means that anti-discrimination law is in constant tension. Following this individualistic legal tradition, anti-discrimination law focuses on recognising the rights of individuals, defining and addressing discrimination as an intersubjective issue, thereby enabling the legal, jurisprudential and doctrinal construction of the right to equality and non-discrimination as an element that is mostly treated as independent and separate from structural discrimination, hence ignoring the origin of inequality and discrimination.[1113] However, the tensions between the individual and collective dimensions of the rights to equality and non-discrimination still endure and are, in many instances, difficult to settle. Hence, since privacy instruments focus on individual rights that do not face the tensions encountered between the collective and individual dimensions of the right to equality, it is easier to articulate a legal framework to prevent algorithmic discrimination from the perspective of privacy, that is, individual data protection.

3.2. THE PROTECTION AND HORIZONTAL EFFECT OF THE FUNDAMENTAL RIGHT TO DATA PROTECTION

The fundamental right to data protection, which originates from the right to privacy,[1114] has been extensively regulated at the EU and member state level for decades. The lengthy development of the right to data protection contained in the GDPR reduces, to a certain

[1112] *Idem*, p. 2102
[1113] RUBIO, A., "Las políticas de igualdad…", *cit.*, 2003, p. 17.
[1114] Spanish Constitutional Court Judgment No. 292/2000, 30th November, section 5.

extent, the gaps that would be generated in the applicability to private parties of the right to data protection if theories that argue for the indirect horizontal effect of fundamental rights were adopted.

The regulatory instruments that develop the right to data protection are mainly aimed to set a minimum threshold of protection for individuals in their interactions with private parties that collect and process their data and make decisions based on said information.[1115] Although the right to data protection also applies and must be respected by public bodies, it has been mainly set up as a safeguard that has to be applied in private interactions, which, as will be discussed later on, produces important deficits in the accountability of public automated decision-making.[1116] In addition, the right to data protection derives from one of the core fundamental rights that the liberal state is constructed upon, the right to privacy. Hence, the discussion regarding the conflicts generated between this right and other fundamental rights and freedoms has not been as intense as discussions concerning the trade-offs between freedom and equality, particularly when equality is approached from substantive perspectives.

There are still certain provisions in the GDPR that are contentious and have generated significant debate, particularly when it comes to the correct interpretation of the transparency requirements contained in the Regulation.[1117] Moreover, the European data protection framework has not gone without criticism by commentators that fear it is stifling innovation.[1118] These views have however only been expressed by a minority of scholars and discussions on data protection have mostly focused on the content of some of the rights recognised to data subjects and how certain provisions should be interpreted rather on criticisms to the data protection framework as a whole.

[1115] BOIX PALOP, A., "Los algoritmos son reglamentos: la necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones", *Revista de Teoría y Método*, vol. 1, 2020, pp. 244-249.
[1116] *Ibidem.*
[1117] BAYAMLIOĞLU, E., "Transparency of automated decisions in the GDPR…", *cit.*, 2018; EDWARDS, L. & VEALE, M., "Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for", *Duke Law & Technology Review*, vol. 16, 2017, pp. 18-84; GOODMAN, B. & FLAXMAN, S., "European Union regulations on algorithmic decision making and a "right to explanation", *AI Magazine*, vol. 38, No 3, 2017, pp. 50-57; KAMINSKI, M. E., "The right to explanation, explained", *Berkeley Technology Law Journal*, vol. 34, No. 1, 2019a, pp. 189-218.
[1118] BOVENBERG, J. *et al*., "How to fix the GDPR's frustration of global biomedical research", *Science*, vol. 370, No. 6512, pp. 40-42.

CHAPTER II. APPLYING THE EU DATA PROTECTION FRAMEWORK TO ALGORITHMS

The right to data protection is recognised in article 8 of the Charter of Fundamental Rights of the EU. It is also indirectly recognised under the right to respect for private and family life included in article 7 of the Charter of Fundamental Rights of the European Union and article 8 of the European Convention on Human Rights. EU secondary law recognises the right to data protection in the GDPR and the Directive for data protection in law enforcement.

This chapter mainly focuses on analysing the GDPR. The GDPR is a comprehensive regulatory instrument that, amongst other objectives, aims to protect individuals from the risks that new data processing technologies pose. However, it is important to keep in mind that, while the focus of this dissertation are algorithms and, hence, the focus of this chapter is the way in which the EU data protection framework applies to algorithms, the legal instruments analysed aim to protect individuals from all forms of data collection and processing.

The Regulation analysed draws protections against the automated processing of personal data from the perspective of individual rights [1119] and, to a certain extent, governance instruments.[1120] The objective of this part is to convey how European regulators have aimed to address many of the issues and concerns regarding data processing technologies by developing the privacy legal framework. Since privacy has been the main perspective from which personal data issues have been traditionally approached, it makes sense to apply that same framework to the new concerns generated by the on-going development of data processing technologies. References to Directive 2016/680 for data are also made throughout this chapter.

Data protection regulatory instruments protect individuals' privacy with a three-legged approach. The first part of this data protection system establishes the nucleus of the anticlassification approach, that is, it prohibits processing certain types of data or altogether prohibits certain types of processing that can be especially risky, such as solely automated processing. The second part or leg provides individuals with a series of rights in order to access and challenge the logic behind automated or semi-automated decision-making

[1119] EDWARDS, L. & VEALE, M., "Slave to the algorithm? …", *cit.*, 2017, p. 74.
[1120] KAMINSKI, M. E., "Binary governance…", *cit.,* 2019b, p. 1595.

processes that concern them. The final part of this system provides governance tools to ensure that those collecting, processing and using automated decision-making respect fundamental rights and other requirements from the onset of automated processes.

In this sense, it is important to highlight the two main streams of research regarding the regulation of algorithms from the privacy perspective. Part of the scholarship has focused on regulating the use of algorithms through the protection of individual rights and transparency, while the other part has addressed other regulatory and technical possibilities more focused on governance.[1121] While this division in the literature is not strict and some scholars have approached algorithmic regulation from both perspectives there is, in general, a clear differentiation.[1122] Both perspectives will be addressed throughout the following chapters.

Most of the scholarship has focused on the first and second parts of this three-legged system and, in particular, on the right to transparency. However, this does not mean that governance instruments that do not rely exclusively on the protection of individual rights have not been developed or are not necessary. For example, within the GDPR there are several provisions that address the way in which algorithms are built and used through forms of auditing and codes of conduct.[1123]

The following sections examine how the privacy approach has been shaped in the EU and analyses how the different trade-offs that result from this particular type of protections against discrimination have been dealt with. As the previous chapter indicated, there are many competing interests that arise when regulating the use of automated or semi-automated decision-making. There are different materialisations of the trade-off between efficiency and equality within the regulation of algorithms from the perspective of individual right protection. They can be distinguished as the trade-off between efficiency and privacy and the trade-off between efficiency and transparency. While the former appears with respect to provisions that are mainly designed to prevent harms to privacy rights from being caused, the second mainly focuses on the recognition of the catalogue of individual rights. Given the relevance that these trade-offs acquire within the privacy framework, this part focuses on explaining said regulatory framework through the conflicts that regulators face.

[1121] *Idem*, pp. 1532-1533.
[1122] *Ibidem.*
[1123] EDWARDS, L. & VEALE, M., "Slave to the algorithm? …", *cit.*, 2017, p. 23; GOODMAN, B., "A step towards accountable algorithms?...", *cit.*, 2016, pp. 4-5.

This chapter is structured following the logic of the three-legged system that was mentioned earlier on. It is important to highlight that, while the main objective is to convey how the EU data protection framework can work as antidiscrimination legislation, the analysis carried out will be comprehensive and not solely focus on the rights to equality and non-discrimination but also, more generally, on how these instruments protect the rights and interests of individuals.

1. THE INFORMATIONAL PRIVACY FRAMEWORK. GENERAL ASPECTS

Before delving into the specific rules that are aimed towards preventing discrimination through anti-classification or that offer individuals and third parties tools to control and detect discrimination in automated or semi-automated decision-making it is important to set the general framework within which the European legal privacy tradition has been developed. This chapter analyses the privacy tradition in the EU, focusing on the scope of application of privacy instruments and the general Fair Information Practice Principles that set the blueprint for privacy legal frameworks.

The origins of data protection regulation in Europe can be found in the Council of Europe's recommendation 509/1968 regarding "Human rights and modern scientific and technological developments" which resulted in the adoption of Convention 108 for the protection of individuals with regard to the processing of personal data. Thus, privacy regulations in Europe in the form of data protection have evolved from the perspective of protecting human dignity and personality rights.

Within the framework of data protection, the discussions and specific data protection regulations that have been passed at the European and national level since the 1970s have aimed towards setting out a framework of protection from data processing technologies built from the dignitary perspective given the fact that the processing of individuals personal data through automated systems is considered to be dehumanising.[1124] Regulatory instruments in Europe have thus focused on comprehensively restricting throughout all sectors data collection, processing and usage, especially when no human intervention is involved, due to concerns regarding the objectification of human beings.[1125]

[1124] JONES, M. L., "The right to a human in the loop…", *cit.*, 2017, p. 221.
[1125] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1541.

The European privacy framework is quite complex seeing as it encompasses both the right to privacy and the right to data protection,[1126] the relationship between which has been debated by the literature at large.[1127] Drawing on LYNSKEY's work, this dissertation approaches the relationship between the rights to privacy and data protection in Europe as complementary rights built upon dignitary concerns[1128] which, although they sometimes protect different elements, in many cases work together.[1129] Thus, the right to data protection should be interpreted from the perspective and as an expression of informational privacy.

This interpretation of the relationship between both rights is in line with the European Court of Justice, which, for example, in paragraph 53 of the Judgment for Joined Cases C-293/12 and C-594/12[1130] established that "the protection of personal data resulting from the explicit obligation laid down in Article 8(1) of the Charter is especially important for the right to respect for private life enshrined in Article 7 of the Charter."

Thus, the data protection regulatory framework developed through the GDPR and the Directive for data protection in law enforcement conveys the complementarity between both rights seeing as individual data protection is structured through prohibitions[1131] to collect and process data which are clear expressions of the right to privacy.

1.1. THE SCOPE OF APPLICATION OF INFORMATIONAL PRIVACY REGULATIONS

Regulations that protect privacy through data collection and processing prohibitions can be designed from different perspectives that can also be combined and which include banning some actors, such as employers, from collecting and/or accessing certain information, restricting the processing of data for certain purposes, restricting certain types of data processing and prohibiting data-holders to disclose or disseminate certain information.[1132] A

[1126] Since, in the US, data protection regulations are drawn from the perspective and as an expression of privacy rights, the relationship between both elements does not generate as many interpretative problems.
[1127] KOKOTT, J., & SOBOTTA, C., "The distinction between privacy and data protection in the jurisprudence of the CJEU and the ECtHR", *International Data Privacy Law*, vol. 3, No. 4, 2013, pp. 222-228; LYNSKEY, O., *The Foundations of EU Data Protection Law*, Oxford, Oxford University Press, 2015, pp. 89-130.
[1128] LYNSKEY, O., *The Foundations of EU Data Protection Law*, *cit.*, 2015, pp. 94-99.
[1129] *Idem*, p. 105.
[1130] CJEU Judgment 8th April 2014, Joined Cases C-293/12 and C-594/12, Digital Rights Ireland Ltd v. Minister for Communications, Marine and Natural Resources and Others and Kärntner Landesregierung and Others.
[1131] Since prohibitions to data collection and processing are general mandates they may seem to fall under the scope of the systemic regulation of algorithmic decision-making or collective governance perspective. However, these prohibitions have been developed in both US and EU legislation within the framework and in order to reinforce individual right protection. KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1589-1591.
[1132] ROBERTS, J. L., "Protecting privacy to prevent discrimination", *cit.*, 2015, pp. 2107-2108.

key element and common trait in Western legal approaches to privacy is the fact that only Personally Identifiable Information (PII), that is, non-anonymised information, is included under the scope of protection of privacy regulations seeing as it is considered that harms cannot be caused to individuals by information that cannot be linked back to them.[1133]

1.1.1. Anonymisation

Anonymised datasets do not fall within the scope of application and resulting restrictions contained in data protection regulatory instruments in the EU. Since individuals cannot be identified from the data contained in anonymised datasets it is considered that they will not be at risk of being harmed by the unfair outcomes, including discrimination, which can result from automated data processing. However, given the level of development of data processing technologies, in order to ensure that this protection is effective, datasets must comply with a series of requirements to be considered anonymised.

The Article 29 Working Party (A29WP)[1134] Opinion 05/2014 on Anonymisation Techniques established three elements that should be controlled for in anonymisation techniques in order to ensure that, once a dataset had been anonymised, the individuals whose information was contained in it could not be re-identified:[1135]

- "Singling out, which corresponds to the possibility to isolate some or all records which identify an individual in the dataset;

- Linkability, which is the ability to link, at least, two records concerning the same data subject[1136] or a group of data subjects (either in the same database or in two different databases). If an attacker can establish (e.g. by means of correlation analysis) that two records are assigned to a same group of individuals but cannot single out individuals

[1133] OHM, P., "Broken promises of privacy: responding to the surprising failure of anonymization", *UCLA Law Review*, vol. 57, 2010, p. 1740.
[1134] The ARTICLE 29 WORKING PARTY (A29WP) was the responsible body for interpreting the DPD and the GDPR until it was substituted by the European Data Protection Board (EDPB), which endorsed the A29WP's guidelines. See EUROPEAN DATA PROTECTION BOARD, "GDPR: Guidelines, recommendations, best practices". Available on 11th May 2019 at: https://edpb.europa.eu/
[1135] ARTICLE 29 WORKING PARTY, "Opinion 05/2014 on Anonymisation Techniques", 0829/14/EN, 10th April 2014, p. 11.
[1136] According to article 4(1) of the GDPR a data subject is a an identifiable natural person, who "is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".

in this group, the technique provides resistance against "singling out" but not against linkability;

- Inference, which is the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes."

The ability of linking different pieces of data belonging to the same individual is sometimes necessary, amongst other fields, for demographic and medical research in order to carry out longitudinal studies, that is, studies that analyse certain characteristics of individuals during the course of several years.[1137] While the ability of linking information pertaining to the same individual contained in different datasets is very risky and can be a first step in the re-identification process, said risk is largely reduced when the information linked belongs to the same dataset because of the limited amount of information a single dataset contains and, in particular, due to the fact that all the data has been anonymised following the same techniques.[1138] In those instances, the risk of re-identification is minimal while the information extracted could be highly beneficial for research and practical application purposes.[1139]

A similar assessment can be made with regard to controlling the inferences made by the data processing tools employed. There is a fundamental difference between attribute and identity disclosure,[1140] while the former can lead to the latter it does not necessarily do so. Additionally, although attribute disclosure through inferences can result in unfair and discriminatory decisions being made, it depends on how the information obtained from the processed data is used.[1141] By limiting the possibility that a model has of obtaining inferences from a dataset, much of the analytic utility of the data can be lost and the socially beneficial results that could be obtained from said model largely reduced.[1142]

1.1.2. Pseudonymisation

Pseudonymisation is the substitution of a piece of information corresponding to the individual such as her name or surname for other information such as her ID or social

[1137] GIL GONZÁLEZ, E., *Big data, Privacidad y Protección de Datos*, *cit.*, 2015, p. 97.
[1138] *Idem*, p. 98.
[1139] *Ibidem*.
[1140] EL EMAM, K. & ÁLVAREZ, C., "A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques", *International Data Privacy Law Journal*, vol. 5, No. 1, 2015, p. 77.
[1141] *Idem*, p. 78.
[1142] *Ibidem.*

security number or even a random code that although makes it more difficult to identify the individual does allow for her indirect identification. Pseudonymisation is thus not an anonymisation technique[1143] and the information contained in the resulting datasets is still considered to be personal data and is therefore included within the scope of application of the GDPR.[1144]

Pseudonymisation does however have very important effects regarding the prevention of algorithmic discrimination seeing as it hinders the identification of an individual's race, gender or other sensitive information. Furthermore, as long as the pseudonym is not shared between different datasets the risk that the individuals whose information is contained in the dataset will be identified is quite low.[1145] Additionally, even if the risk of identification increases, if the relevant elements are controlled for, pseudonymisation can be used in order to establish relationships between different pieces of data belonging to the same individual that can be very useful for research without said individual being identified.[1146]

Consequently, once again, the positive or negative consequences that may result from data processing largely depend on how machine learning algorithms are employed in the processing of pseudonymised data. If the correct protections to avoid identification were put in place it would not be necessary to restrict the processing of pseudonymised datasets or to establish such a restrictive interpretation of what is considered an anonymised dataset under EU regulations. However, since new technological developments allow facilitate the identification of individuals, the European legislator decided to tilt the scales in favour of extensive privacy protections.

1.1.3. Scope of application of the EU's data protection framework

The EU has adopted a general approach to informational privacy which thus covers all types of situations in which personal data collection and/or processing is involved. It is therefore convenient to specifically analyse the scope of application of the EU data protection regime.

[1143] *Idem*, p. 90.
[1144] Recital 26 of the GDPR.
[1145] GIL GONZÁLEZ, E., *Big data, Privacidad y Protección de Datos*, *cit.*, p. 91.
[1146] INFORMATION COMMISSIONER'S OFFICE, "Anonymisation: managing data protection risk code of practice", 2012, p. 21: "…even though pseudonymised data does not identify an individual, in the hands of those who do not have access to the 'key', the possibility of linking several anonymised datasets to the same individual can be a precursor to identification. This does not mean though, that effective anonymisation through pseudonymisation becomes impossible. The Information Commissioner recognises that some forms of research, for example longitudinal studies, can only take place where different pieces of data can be linked reliably to the same individual."

In order to make sure that all cases of data collection and processing are covered, the EU framework contains two regulatory instruments that complement each other. The GDPR is applicable in all situations except in cases falling under the scope of the Directive for data protection in law enforcement and the criminal justice system. Hence, the GDPR will be applied when data collection and/or processing is not carried out for purposes regarding the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties.

Moreover, following the European privacy and data protection tradition that aims towards creating a regulatory framework that broadly covers and protects individuals against as many risks derived from the collection and processing of their data as possible, the GDPR defines "personal data" in very extensive terms. Therefore, although the Regulation only applies to the protection of personal data, since this type of data defined as "any information relating to an identified or identifiable natural person",[1147] any information that is not anonymised is covered by the GDPR, even pseudonymised data.[1148]

The territorial scope of application of the Data Protection Regulation (article 3) also illustrates the legislator's will to offer as broad a protection as possible given the fact that the GDPR applies to the processing of data even when it is not carried out within the EU as long as said processing takes place "in the context of the activities of an establishment of a controller or a processor in the Union". Furthermore, even when a controller or processor is not established in the EU the GDPR applies when "processing activities are related to the offering of goods or services" or to the monitoring of data subjects' "behaviour as far as their behaviour takes place within the Union". Finally, the Regulation is also applicable "to the processing of personal data by a controller not established in the Union, but in a place where member state law applies by virtue of public international law".

1.2. PRIVACY PRINCIPLES

A very important element in data protection regulations in Western privacy frameworks is the use of Fair Information Practice Principles (FIPPs) or Fair Information Practices (FIPs) as the base upon which privacy regulations are built.[1149] These principles originated in the US but

[1147] See trade-offs regarding anonymised and pseudonumised datasets in section I of this Chapter.
[1148] Article 2.2(d) of the GDPR.
[1149] CATE, F. H., "The failure of fair information practice principles", in WINN, J. K., *Consumer Protection in the Age of the Information Economy*, Abingdon, Routledge, 2006, p. 341.

were adopted by the OECD and have thus been widely used to conform European data protection regulations[1150] seeing as they include providing individuals with information regarding the collection and processing of their data; the need for individuals' consent regarding the collection, processing, usage or disclosure of their information; providing individuals with the possibility of accessing collected and processed data and contesting any inaccuracies; ensuring that data is accurate and secure and setting up enforcement mechanisms to ensure that all the other principles are correctly enforced.[1151]

The following sections analyse the privacy principles included in the GDPR seeing as it is the most comprehensive privacy regime and includes principles applicable to data collection, processing, usage, quality and storage[1152] whereas most other privacy regimes focus on setting data processing principles and not as much on the other aspects.[1153] Article 5 of the GDPR established the principles that apply to the processing of personal data, which specifically are: 1) lawfulness, fairness and transparency; 2) purpose limitation; 3) data minimisation; 4) accuracy; 5) storage limitation; and, 6) integrity and confidentiality.

1.2.1. Data processing principles: lawfulness, fairness, transparency, integrity and confidentiality

The principles recognised in paragraphs 1) and 6) of article 5 are analysed together as they all refer to requirements that apply in data processing.

The exact definition of what lawful processing of personal data means is contained in article 6 of the GPDR which establishes the conditions in which said processing is considered to be lawful and also provides member states of the EU with the possibility of adapting some of the lawful bases for processing. Situations in which the processing of personal data is to be considered lawful generally fall under two categories: 1) consent of the data subject and 2) where processing is necessary for a series of purposes that will be detailed later on.

Lawful processing on the basis of consent granted by the data subject is the first condition included in article 6, which conveys how the GDPR is articulated around the idea of informational self-determination and thus places the individual as the core element in the

[1150] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1585.
[1151] CATE, F. H., "The failure of fair information practice principles", *cit.*, 2006, p. 341.
[1152] *Idem*, p. 356.
[1153] HARTZOG, W., "The inadequate, invaluable fair information practices", *Maryland Law Review*, vol. 76, No. 4, 2017, p. 980.

regulation by providing her with individual rights and the power to decide how her information should be collected and processed and to withdraw consent whenever she deems appropriate. The importance of consent must not be overlooked as there are certain types of processing for which none of the other lawful bases can apply, meaning controllers will have to rely on the data subject's consent in order to process her information.[1154]

Nonetheless, and while consent does play a very important role throughout the whole Regulation, there are five other possibilities that controllers can resort to in order to ensure that their data processing activities are lawful. Processing is considered lawful when it is necessary "for the performance of a contract" or "for the purposes of the legitimate interests pursued by the controller or by a third party", which undoubtedly grants controllers with a wide margin of appreciation and action. Legitimate interests are however limited and "overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data". This is particularly important towards the purposes of this dissertation as it reinforces the idea that the protection of the right to equality and non-discrimination is built from the perspective of anti-classification in the GDPR, seeing as only if fundamental rights require the protection of personal data will they override the legitimate interests of the controller or a third party.

Article 5 of the GDPR also establishes that processing must be fair, which is broadly interpreted to mean that data should not be processed in ways that are not expected by data subjects and is therefore closely related to purpose limitation and specification.[1155] The European Data Protection Board has also established that, fair processing of personal data entails "considering possible adverse consequences processing may have on [data subjects], and having regard to the relationship and potential effects of imbalance between them and the controller".[1156]

The broad definitions or indications of what fairness constitutes convey the difficulties of conceptualising fairness that result both from the way in which this notion appears in the

[1154] ZUIDERVEEN BORGESIUS, F., & POORT, J., "Online price discrimination and EU data privacy law," *Journal of Consumer Policy*, vol. 40, 2017, pp. 359-362.
[1155] EUROPEAN DATA PROTECTION BOARD, "Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects", 8th October 2019, p. 6; INFORMATION COMMISSIONER'S OFFICE, "Guide to the General Data Protection Regulation", 2018, pp. 19-20. Available on May 17th 2019 at: https://ico.org.uk/
[1156] EUROPEAN DATA PROTECTION BOARD, "Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects", *cit.*, 2019, p. 6.

GDPR and also due to the general lack of concretion of this concept. While "lawfulness" and "transparency" are developed and detailed in several provisions throughout the GDPR (and, even then, there has been a lengthy debate regarding on what level of "transparency" is demanded in the GDPR), "fairness" only appears as a descriptive element that is added to other types of requirements.[1157]

It is important to highlight that the very concept of fairness has been largely debated in relation to algorithmic processing and, especially, with regard to algorithmic discrimination in recidivism risk prediction.[1158] These conflicts that have arisen in the scholarship convey the difficulties of determining what "fair processing" of personal data constitutes and the need to establish fairness on a contextual basis.[1159] This does not mean that fairness should be determined on a case-by-case basis but that more specific rules should be passed indicating in which sectors and for which purposes a stricter notion of fairness is required and what said notion of fairness entails. The greater degree of specification of the concept of fairness is necessary especially with regard to establishing the degree of protection that is awarded to traditionally disadvantaged groups, for fairness should determine the extent to which a type of processing that results in the perpetuation of said disadvantage can be considered fair, and therefore lawful, under the GDPR.

The transparency principle is specified in a series of information, access and explanation rights contained and developed through the text of the GDPR. Transparency is essential in protecting the right to equality and non-discrimination for it provides data subjects and also possibly public and private oversight parties with the possibility of controlling that algorithms are not processing personal data in a discriminatory manner or yielding discriminatory results. Moreover, in those cases in which results are not discriminatory but automated processing is used for decisions that affect particularly vulnerable populations, for example when adjudicating public aid, transparency is necessary in order to ensure that applicants' data is being processed correctly.[1160]

[1157] WACHTER, S. & MITTELSTADT, B. D., "A right to reasonable inferences: re-thinking data protection law in the age of big data and AI", *Columbia Business Law Review*, vol. 2019, No. 2, 2019, pp. 582.

[1158] CHOULDENOVA, A., "Fair prediction with disparate impact…", *cit.*, 2016; DRESSEL, J. & FARID, H., "The accuracy, fairness, and limits of predicting recidivism", *cit.*, 2018; KLEINBERG, J., MULLAINATHAN, S., & RAGHAVAN, M., "Inherent trade-offs in the fair determination of risk scores", *cit.*, 2016.

[1159] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, p. 10.

[1160] See, for instance, the cases addressed in BELMONTE, E., "La aplicación del bono social del Gobierno niega la ayuda a personas que tienen derecho a ella", *cit.*, 2019 and EUBANKS, V., *Automating Inequality…*, *cit.*, 2017.

Finally, the GDPR establishes two data security[1161] related principles that require the "integrity and confidentiality" of the personal data that is processed. The data security principles require firms to adopt the necessary measures to ensure that the data is not compromised and entail, amongst other obligations, that only authorised individuals can access, modify or disclose the data being processed.[1162] The security principles are developed in article 32 of the GDPR, which establishes the rules that controllers and processors must follow in order to ensure the security of personal data.

1.2.2. Data collection principle: purpose limitation

Article 5(b) establishes that:

"[data shall be] collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes…"

As it was previously discussed, this principle is closely related to the "fair processing" mandate as data subjects must be informed of the purposes for which their data will be processed and processing is limited to the purposes of which the subject has been informed. Part of the literature has argued that this principle may pose problems due to the fact that one of the elements that characterises automated processing and the new technologies involved in said activities, such as machine learning, is the fact that controllers and processors are sometimes not even aware of what the tools used will do with the data or the types of results that they will yield.[1163]

This does not seem to be a valid argument due to the fact that, even if the ways in which self-learning tools process personal data cannot be fully foreseen in advance, the purposes for which the results will be used by the controller will be pre-established. What this principle simply aims to achieve is avoiding that data are generally collected and processed with no

---

In said instances, the algorithms used did not yield discriminatory results but are used in services that are mainly addressed towards vulnerable populations and in which, mistakes in processing lead to perpetuating the disadvantage and oppression of members of said groups.

[1161] ESKENS, S. E., "Profiling the European consumer in the internet of things: how will the General Data Protection Regulation apply to this form of personal data processing, and how should it?", 22nd March 2016, p. 32. Available on 19th March 2020 at: https://papers.ssrn.com/; INFORMATION COMMISSIONER'S OFFICE, "Guide to the General Data Protection Regulation", 2018, p. 209.

[1162] INFORMATION COMMISSIONER'S OFFICE, "Guide to the General Data Protection Regulation", 2018, pp. 209-210.

[1163] ZARSKY, T., "Incompatible: the GDPR in the age of big data", *Seton Hall Review*, vol. 47, 2017, p. 1005-1006.

limitations whatsoever for this would mean heavily curtailing individuals' rights and guarantees when subjected to automated data processing.

In addition, the Regulation allows compatible further processing, which provides controllers with a wide range of further processing possibilities. Only if further processing concerns a purpose that is completely different from the first purpose for which the data was processed will it be considered that the purpose limitation principle has been breached.[1164] Compatibility is to be determined on a case-by-case basis, which undeniably provides controllers with a significant amount of power and flexibility when determining whether a certain purpose is compatible with the initial purpose for which the data was collected and processed.[1165] Nonetheless, the importance that this principle has in controlling the activities of data controllers must not be downplayed, for it does at least set a series of minimum requirements that must be complied with and forces controllers to set limits to the processing activities they carry out in order to safeguard the rights and interests of data subjects.

1.2.3. Data and storage requirements: data minimisation, accuracy and storage limitation

Data minimisation, accuracy and storage limitation principles are analysed together due to the fact that they are all closely linked to one another seeing as they all refer to specific requirements that either the data itself must comply with or that are applied to its storage.

The data minimisation principle forces controllers to only process the data that is relevant towards the stated purposes. This principle is therefore closely related to purpose limitation as the requirements that controllers must comply with in relation to data minimisation will be set by the purpose they specified. This principle is in obvious tension with current data processing technologies seeing as tools such as machine learning algorithms yield better results when they process larger amounts of data.[1166]

The mandates that controllers must follow in order to comply with the specifications of the data minimisation principle must be put in relation with the idea of justifiability,

[1164] INFORMATION COMMISSIONER'S OFFICE, "Guide to the General Data Protection Regulation", 2018, pp. 21-24.
[1165] ARTICLE 29 WORKING PARTY, "Opinion 03/2013 on purpose limitation", *cit.*, 2013, p. 39.
[1166] ZARSKY, T. Z, "Incompatible…", *cit.*, 2017, p. 1010-1011.

explainability and intuitiveness of automated decision-making.[1167] Legal systems are created and structured by the human brain and therefore require many types of decisions to be properly justified in a manner that humans will understand. It would not be logical for a human being that a bank used data on what types of shoes people wore in order to determine loan decisions. Consequently, if controllers have to stick to the requirements set by the data minimisation principle, if a type of data input does not seem logical in relation to the purposes that the controller is processing the data for, it might be considered not to comply with this principle. However, considering that controllers generally tend to formulate processing purposes in a very generic manner and that it is very difficult to control all the data that actually serves as input in automated systems, does not seem to present many limits to the activities of controllers.[1168]

Recital 39 of the GDPR also puts the data minimisation principle in relation to storage limitation as it indicates that data minimisation "requires, in particular, ensuring that the period for which the personal data are stored is limited to a strict minimum". The storage limitation principle can therefore be treated as an expression of data minimisation for the latter, as article 5 establishes it indicating that "personal data shall be […] kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed".

The data minimisation principle can also be analysed with regard to the accuracy principle seeing as they both set requirements for a certain degree of input data quality in the sense that the data minimisation principle, for example, requires the data to be "adequate … to what is necessary for the purposes of the processing". The accuracy principle is specified through some of the rights that the GDPR recognises to individuals, such as the rights to rectification and erasure (right to be forgotten).

The principles studied in this section (data minimisation, accuracy and storage limitation) as well as purpose limitation only seem to apply to input data.[1169] This is particularly relevant since many instances of algorithmic discrimination result from erroneous and inaccurate output data.

[1167] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, pp. 1085-1139.
[1168] WACHTER, S. & MITTELSTADT, B. D., "A right to reasonable inferences…", *cit.*, 2019, p. 583.
[1169] *Ibidem.*

Most existing regulatory systems aimed towards addressing automated or semi-automated decision-making are based on FIPs or FIPPs. As the following chapters will convey, the heavy reliance on individual rights combined with mostly soft-law governance mechanisms hinders the possibility of effectively rendering algorithms accountable through the European data protection framework. Particularly when it comes to the impact of automated systems on the rights to equality and non-discrimination, the narrow focus of data protection instruments and their inability to implement effective accountability mechanisms that could help prevent not just harms to data protection, but also to other rights, including the rights to equality and non-discrimination, highlight the shortcomings of FIP-based regulations. This does not mean that said principles are not relevant or valuable, but that it is not possible to solely rely on them and that, relying on the useful blueprint that they provide, regulators should build a more solid and effective system of protection against the risks generated by algorithmic decision-making.

2. PROHIBITIONS TO ACCESS AND PROCESS INFORMATION

This section mainly focuses on the degree to which data controllers and processors and society as a whole should give up some of the benefits brought by current computing capabilities in order to favour individuals' privacy. The larger the amount of information available to controllers, the more accurate results they will obtain. This accuracy can lead to great advances in research that can benefit both the public and private sector. However, processing certain types of information can result in outcomes that reinforce existing oppressive social structures and lead to specific discriminatory results.

The European data protection framework establishes general prohibitions with regard to the processing of special categories of personal data as well as with regard to solely automated processing. The exemptions to the prohibitions contained in articles 9 and 22 of the GDPR (and 10 and 11 of Directive 2016/680 for data protection in law enforcement) convey how, in certain cases, the risk of obtaining unfair and discriminatory outcomes from data processing can be accepted in exchange for the benefits that automated or semi-automated data processing can bring by.

The following pages carry out an analysis of the main prohibitions contained in the GDPR aimed towards preventing discrimination and other forms of unfair treatment resulting from the processing of personal data will be analysed. Finally, the relevant provisions in Directive

2016/680 for data protection in law enforcement will also be analysed, paying special attention to those elements that differ from the GDPR.

2.1. PRIVACY AS ANTI-DISCRIMINATION THROUGH GENERAL PROHIBITIONS IN THE GDPR

2.1.1. Processing special categories of data

Article 9.1 of the GDPR prohibits the "processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation".

The GDPR also sets up a series of exceptions that would allow for the sensitive data above listed to be processed by data controllers[1170] or processors.[1171] These exceptions generally encompass cases in which the data subject explicitly authorises the use of said information, has made said information public or processing it is necessary for public interest reasons or for the interest of the data subject.

The special categories of personal data considered in article 9 of the GDPR must be understood with regard to the anti-discrimination concerns[1172] expressed in Recital 71 of the GDPR and which include discrimination "inter alia, […] on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation". The objective of article 9 of the GDPR is thus to establish a set of elements that could lead to the personal identification of individuals and that are especially sensitive, seeing as they could provide information on protected group membership consequently enabling situations of individual discrimination.

[1170] GDPR, Art. 4, Definitions: "(7)'controller' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or member state law, the controller or the specific criteria for its nomination may be provided for by Union or member state law".
[1171] GDPR, Art. 4, Definitions: "(8) 'processor' means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller".
[1172] HORDERN, V. "How do you solve a problem like special categories of data?", *Data Protection Leader*, March 2018, p. 6.

The types of data for which a processing prohibition is established in article 9 of the GDPR are labelled "special categories of personal data". This is highly relevant given the fact that, according to article 4 of the GDPR personal data is:

> "…any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person…"

Consequently, the categories of information included under article 9 necessarily relate to elements that can lead to the identification of an individual.

There are two other elements that must be emphasised. Firstly, financial information (and thus socioeconomic status) is not included as a special category of personal data protected under article 9 of the GDPR. However, economic situation is included within the scope of profiling (article 6.4 of the GDPR), which is prohibited by article 22. In addition, in article 4, "social identity" is included as a type of information that can be considered personal data. Nevertheless, excluding financial information as a special category of data means that fewer protections are set for discrimination based on social origin and property status, both of which are protected by article 14 of the European Convention on Human Rights and fall under the construction of classism as a narrative of oppression regarding socioeconomic status.

Additionally, although sex (or gender) and sexual identity could be understood to be protected under sex life, genetic and biometric data under the special categories of personal data listed in article 9, they are certainly not included amongst the explicit antidiscrimination objectives set out in Recital 71 of the GDPR. The lack of inclusion of sex as a special category of personal data is logical seeing as special categories of data convey information that could lead to the identification of the individual whereas being a biologically assigned male or female hardly provides any information that could personally identify an individual and, were it to provide relevant information to identify an individual it would probably have to be combined with other data falling under the categories of sex life and/or genetic and

biometric data. Additionally, the prevention of gender discrimination must be understood to fall under the open clause set up by the phrase "*inter alia*".

However, the fact that gender is not explicitly included in Recital 71 of the GDPR comes to show the extent to which this regulation is structured from the perspective of dominant narratives. Discrimination on the basis of gender is one of the most pervasive and widespread forms of oppression, which affects half of the world's population and which has been repeatedly proven to be embedded in many algorithms and the sector in charge of creating them. Excluding the explicit mention of gender throughout the whole text of the GDPR does nothing but hide gender hierarchy structures under the apparent neutrality of a regulation drawn form the perspective of the prototypical liberal individual.

2.1.1.1. Scope of the prohibition

i) *Personal scope of application: search engine operators*

When analysing the scope of the prohibition set by article 9, the most relevant element is determining whether the prohibition also affects the processing of variables that can act as proxies for special categories of information. However, before delving into the different elements and problems that can arise from the possibility of indirectly processing special categories of information, it is relevant to briefly refer to the personal scope of application of said prohibition when it comes to the activities of search engine operators.

The CJEU ruled in case C-136/17[1173] that the prohibition of processing special categories of personal data does apply to search engine operators, which is highly relevant considering how search engines reinforce dominant narratives through the results shown when certain combinations of words are used in searches. The ruling also referred to the requests made by several individuals to Google in exercise of their right to be forgotten so that the search engine de-referenced a series of results that appeared upon searches of the plaintiffs' names and which were linked to third-party websites that contained information on those individuals, including references to special categories of data. The Court stated that, while said requests must generally be accepted, they must be balanced with the right to information,

[1173] CJEU Judgment 24th September 2019, C-136/17, GC and Others v Commission nationale de l'informatique et des libertés (CNIL).

which would, for instance, outweigh the applicants' right to data protection if the data subject played a significant role in public life.[1174]

ii) *Material scope of application: the proxy problem*

The scope of application of the prohibition depends on whether it is understood that this prohibition only affects variables that directly reveal special categories of personal data or also proxy variables for special categories of data.[1175] The problem with a restrictive interpretation of such a provision is that the discrimination generated by automated decision-making is, in many cases the product of proxy variables for specially protected group membership.[1176] Consequently, exclusively removing the explicit on suspect categories, such as race or gender, renders this particular provision almost void of content seeing as disadvantaged group membership can also be brought into the decision-making process through proxy variables that are correlated with special categories of personal data.[1177]

However, if the more extensive interpretation means that all proxy variables for special categories of personal data cannot be used in automated decision-making this may mean that relevant information is being lost. Moreover, it is very difficult to completely eliminate all proxy variables as it is quite likely that the remaining variables have some sort of correlation to the special categories.[1178]

Thus, a restrictive interpretation of article 9 makes it ineffective while an extensive interpretation is probably unfeasible. Considering this scenario, GOODMAN has argued for the possibility of taking a step backwards and allowing for the introduction of variables that explicitly convey information on special categories of personal data into algorithms.[1179] Although this might seem counterintuitive, the truth is that the elimination of said variables hinders the detection of other proxy variables correlated to special categories of personal data thus making it practically impossible to detect the discrimination produced by a certain

[1174] CJEU Judgment 24th September 2019, C-136/17, GC and Others v Commission nationale de l'informatique et des libertés (CNIL), paragraph 53, in reference to CJEU Judgment 13th May 2014, C- 131/12 Google Spain SL and Google Inc. v. Agencia Española de Protección de Datos (AEPD) and Mario Costeja González, paragraph 99.
[1175] MALGIERI, G. & COMANDÉ, G., "Sensitive-by-distance…", *cit.*, 2017a, p. 233.
[1176] GOODMAN, B., "A step towards accountable algorithms?...", *cit.*, 2016, p. 3.
[1177] *Ibidem.*
[1178] *Ibid.*
[1179] *Ibid.*

automated decision.[1180] Moreover, as several studies have shown, eliminating sensitive information can sometimes lead to wrong inferences that result in especially harmful results for members of protected groups.[1181] This issue will be further addressed in the following chapter.

In any case, focusing on the actual text of the GDPR and the way in which the prohibition contained in article 9 should be interpreted, given the fact that the article prohibits the "processing of personal data revealing" special categories of personal data such as race or religious beliefs, said provision must undoubtedly be interpreted to encompass all data that, when processed, could lead to inferring an individual's special categories of personal data. This is the way in which Article 29 Working Party (A29WP) interpreted the corresponding article in the DPD.[1182]

iii) *Solutions for the discrimination by proxy problem*

However, even if an extensive interpretation of article 9 is adopted, the problem of knowing what data falls under the prohibition remains seeing as, given the current state of the technology and constant developments, it is possible to argue that virtually all data is sensitive data. In order to at least provide some solutions to this problem, MALGIERI and COMANDÉ present a "sensitive-by-distance" approach that aims to detect what data is more susceptible of revealing special categories of information.[1183] They place special emphasis on the fact that privacy regulations should cover all data from which special categories could be directly or indirectly revealed. While they focus on health and quasi-health data, the clear intention is to also apply this extensive interpretation of article 9 of the GDPR to other special categories of information.

The sensitive-by-distance approach provides a way in which to delimit the boundaries between sensitive and non-sensitive data by combining two variables: "intrinsic

[1180] *Ibid.*
[1181] DOLEAC, J. L. & HANSEN, B., "The unintended consequences of 'ban the box'…", *cit.*, 2020; HOLZER, H. J., RAPHAEL, S., & STOLL, M. A., "Perceived criminality, criminal background checks, and the racial hiring practices of employers", *Journal of Law and Economics*, vol. 49, No. 2, 2006, pp. 451-480; STRAHILEVITZ, L., "Privacy versus antidiscrimination", *cit.*, 2008, pp. 363-381.
[1182] ARTICLE 29 WORKING PARTY, "Advice paper on special categories of data ('sensitive data')", 20th April 2011: "The term "data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership" is to be understood that not only data which by its nature contains sensitive information is covered by this provision, but also data from which sensitive information with regard to an individual can be concluded".
[1183] MALGIERI, G. & COMANDÉ, G., "Sensitive-by-distance…", *cit.*, 2017a, p. 238.

sensitiveness" of the data and "computational capacity". The first variable detects whether a piece of information directly reveals special category personal data. For example, "data about health status (e.g. diagnosis, blood pressure, blood counts, etc.), are inherently 'sensitive', whereas data about food consumption (and also daily exercise, air pollution of one's city, etc.) are not intrinsically ' sensitive' , but might be considered sensitive because the computational distance between such data and sensitive data resulting from them is relatively small."[1184]

The way in which the second variable (computational capacity) is measured is through subjective and objective parameters. Measurement of the subjective parameters is obtained through the criteria established in Recital 26 of the GDPR, which entails evaluating the means that will reasonably likely be used when processing information, which are to be calculated by considering "the costs of and the amount of time required" as well as "the available technology at the time of the processing and technological developments". The subjective parameter thus involves measuring the amount of human and economic resources that the data controller will likely devote to a specific data processing task and, consequently, the resulting processing power.[1185] The objective parameters include the personal and non-personal accessory data available as well as the legal rules for data collection, use and reuse.[1186]

The problem with this proposal is the fact that, due to its specificity, it would entail a case-by-case approach. Nonetheless, a variant of this proposal could be introduced in regulations applicable to specific sectors such as the health insurance sector that could be complementary to the broader privacy protection regulatory instruments. By calculating the average measure of the information available to health insurers and the resources they are willing to deploy to process said information it would be possible to establish limitations on the information that could be collected. While it is true that the rapid development of data processing technologies would make it necessary for these regulations to be constantly updated, they would at least provide a more specific rule of what is considered sensitive information under the special categories of article 9 of the GDPR and would provide better safeguards against the processing of information that could lead to the discrimination of members of traditionally oppressed groups.

[1184] *Ibidem*.
[1185] *Ibid.*
[1186] *Ibid.*

There are also several proposals put forward by computer science scholars on how to minimise the introduction of sensitive attributes when creating algorithms.[1187] The objective of these proposals is to eliminate sensitive attribute inputs while maintaining the same level of predictive accuracy.[1188] However, while these approaches manage to keep an adequate level of algorithmic predictive accuracy, they fail to eliminate all sensitive information by association from the algorithm and render automated systems inapplicable in cases in which said sensitive attributes are very relevant for the tasks being conducted.[1189]

In the current scenario, in which there is no comprehensive oversight mechanisms through which to detect instances of algorithmic discrimination and in which these cases and, in general, cases in which harms are caused to the rights to data protection are only being detected practically by chance,[1190] the prohibition set by article 9 of the GDPR sets an necessary safeguard in the processing of data that can cause specially significant harms to individuals. However, as other regulatory instruments for the control of algorithms are developed, it may be necessary to rethink this prohibition as it can sometimes hinder the possibility of detecting algorithmic discrimination.

2.1.1.2. Processing of personal data relating to criminal convictions and offences

Article 10 of the GDPR only allows for criminal conviction and offence data to be processed either "under the control of official authority" or as long as "appropriate safeguards for the rights and freedoms of data subjects". While the way in which the rule is expressed varies from article 9 in that it does not include an explicit prohibition, the general rule of the provision clearly forbids the processing of said information unless one of the two aforementioned exceptions takes place. An additional limitation is also established with regard to the possession of data on criminal convictions which has to be "kept under the control of official authority" with no exceptions. These safeguards, essential to prevent discrimination produced as a consequence of the use of algorithms in law enforcement, are

[1187] EDWARDS, H. & STORKEY, A., "Censoring representations with an adversary", 2015, pp. 1-14. Available on 8th February 2020 at: https://arxiv.org/; LOUIZOS, C. *et al.*, "The variational fair autoencoder", 2015. Available on 8th February 2020 at: https://arxiv.org/.
[1188] CRIADO, N. & SUCH, J. M., "Digital discrimination", in in YEUNG, K. & LODGE, M., (eds.), *Algorithmic Regulation*, Oxford, Oxford University Press, 2019, p. 90.
[1189] *Ibidem*.
[1190] See, for example, EUROPEAN DATA PROTECTION BOARD, "Hamburg commissioner fines H&M 35.3 million euro for data protection violations in service centre", 2020. Available on 3rd October 2020 at: https://edpb.europa.eu/

also reinforced by Directive 2016/680 for data protection in law enforcement, which will be addressed later on.

2.1.2. The right (or general prohibition) not be subject to decisions based solely on automated processing, including profiling

2.1.2.1. The right not to be subject to a decision based solely on automated processing, including profiling

Article 22 of the GDPR specifically focuses on automated decision-making. Said provision grants data subjects the right "not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her" and establishes specific safeguards for those cases regarding the special categories of data referred to in article 9.1 that will later be discussed. It is important to highlight that the right not to be subject to automated decisions is built as a general prohibition that can only be lifted if one of the exceptions applies rather than as a right that the data subject has to request not to be subjected to automated decision-making.[1191]

This is highly relevant given the fact that, even though this provision is included under the general catalogue of individual rights contained in articles 12 to 22, the way in which it is set out brings it closer to a general rule, or rather prohibition, for processing data than to a right that can be exercised by individuals. The A29WP confirmed this interpretation.[1192]

With respect to profiling, it is specifically defined in article 4 of the GDPR as:

> "…any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements".

[1191] BYGRAVE, L. A., "EU data protection law falls short as desirable model for algorithmic regulation", *Centre for Analysis Risk and Regulation Discussion Papers*, No. 85, 2017, p. 33.
[1192] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", 17/EN, WP 251rev.01, 6th February 2018, p. 19: "Article 22(1) establishes a general prohibition for decision-making based solely on automated processing. This prohibition applies whether or not the data subject takes an action regarding the processing of their personal data".

As it was already discussed, drawing up profiles on individuals can to a certain extent be interpreted as a form of automated decision-making and it increases the risks of discrimination.[1193] In many cases, such as in recruiting or lending decisions, profiling will take place alongside automated decision-making. However, in other instances, such as when systems that automatically issue traffic fines are used, automated decision-making without profiling is used.[1194] More importantly, automated profiling can be carried out without an actual decision being made regarding the individual in which case, although profiling can be argued to be in itself a form of automated decision-making process, it would be much more difficult to defend this form of processing to be comprehended under the prohibition of article 22. The specific inclusion of both types of processing in the general prohibition thus offers a more comprehensive protection of individual rights and against cases of algorithmic discrimination.

2.1.2.2. Exceptions to the right recognised in article 22 and safeguards

Paragraph 2 of article 22 establishes a series of exceptional situations in which the prohibition does not apply:

> "(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
>
> (b) is authorised by Union or member state law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
>
> (c) is based on the data subject's explicit consent."

It is important to highlight that, in the same way that exception (b) listed above indicates that the EU or member state whose law authorises decisions based solely on automated processing must lay down "suitable measures to safeguard the data subject's rights and freedoms and legitimate interests", paragraph 3 of article 22 establishes a similar mandate for exceptions (a) and (c). Paragraph 3 does, however, add on a minimum requirement to these safeguard measures, indicating that they must at least consist of "the right to obtain human

[1193] HARDT, M., "How big data is unfair", *Medium*, 26th September 2014. Available on 6th May 2019 at: https://medium.com/
[1194] BAYAMLIOĞLU, E., "Transparency of automated decisions in the GDPR...", *cit.*, 2018, p. 25.

intervention on the part of the controller, to express his or her point of view and to contest the decision".

The fact that a minimum requirement for the safeguard measures is set for those cases in which the exception is not authorised by EU or member state law makes absolute sense given that the exceptions regarding the entering of a contract and the data subject's explicit consent pose greater risks on the data subject. The many procedures a legislative act must go through in order to be passed both at the national level and at the European level, as well as the possibility that citizens have to participate in the legislative process and the many existing mechanisms that may be employed with the objective of challenging laws and other types of rules means that it offers the subjects to whom it applies greater guarantees that their rights, freedoms and legitimate interests will be correctly protected and that the safeguards put in place will adapt to the processes that they apply to.[1195]

However, even when no specific safeguard mechanisms are put in place by member states, the provision should be subjected to an extensive interpretation and also include the minimum safeguards required for exceptions (a) and (c) of article 22.2 of the GDPR. This requirement is derived from Recital 71, which states the following:

> "… [D]ecision-making based on such processing, including profiling, should be allowed where expressly authorised by Union or Member State law to which the controller is subject … such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision."[1196]

The Recital, while not a rule in itself,[1197] clearly underscores the minimal required safeguards that are to be put in place when the exceptions to article 22.1 or 22.4 apply, which means that if no specific safeguards are implemented by member states, at least those included in article 22.3 of the GDPR should be considered applicable.

[1195] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, p. 246.
[1196] The right to explanation and other associated rights will be addressed throughout this chapter.
[1197] CJEU Judgment 13th July 1989, C-215/88, Casa Fleischhandel v. Bundesanstalt für landwirtschaftliche marktordnung.

It is important to add that the three rights or safeguards set by article 22.3 of the GDPR actually constitute active rights of individuals and will thus be analysed later on in the section dedicated to due process rights contained in the EU data protection framework.

2.1.2.3. Special protections for decisions based solely on the automated processing of special categories of personal data

Finally, paragraph 4 of article 22 limits even further the possibility of carrying out decisions based solely on automated processing in those cases in which the decision is "based on special categories of personal data referred to in article 9(1)" of the GDPR, that is, "personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation".

For the cases in which the decisions are based on special categories of personal data, not only must one of the exceptions in article 22.2 be present, but one of the two following cases regulated in article 9.2 (a) and (g) must also apply:

> "(a) The data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or member state law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject.
>
> [...]
>
> (g) Processing is necessary for reasons of substantial public interest, on the basis of Union or member state law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject."

In these cases it does not suffice with the general consent that is required in article 6(1)(a) of the GDPR for the processing of personal data to be considered lawful. In order for the exception to the prohibition to process special categories of personal data to apply it is

necessary that the consent provided by the individual specifically refers to the processing of said categories of personal data.[1198]

It is important to highlight that, the special protection offered by article 22.4 can be considered to also cover gender through the special categories of personal data revealing sex life, genetic and biometric data, for example. However, said intensified protection is not applied to economic status for there are no special categories of data through which it may be inferred to be comprehended within the special categories of data of article 9.

2.1.2.4. Issues raised with regard to the scope of article 22.1

i) *Decisions based solely on automated processing*

Article 22 limits its prohibition (or right not to be subject) to decisions "based solely on automated processing". Consequently, part of the literature understood that any minimal and insignificant human involvement would suffice to render article 22 and its related provisions inapplicable.[1199] This conclusion makes absolute sense from the perspective of a literal interpretation of the article and also if we consider one of the amendments to the original text that was not finally included in the Regulation.

The European Parliament proposed an amendment to article 20 of the original text (article 22 in the final version) so that it read as follows:

> "Profiling which leads to measures producing legal effects concerning the data subject or does similarly significantly affect the interests, rights or freedoms of the concerned data subject shall not be based solely or predominantly on automated processing and shall include human assessment…"

[1198] DRECHSLER, L. & BENITO SÁNCHEZ, J. C., "The price is (not) right…", *cit.*, 2018, p. 8; CJEU Judgment 24th September 2019, C-136/17, GC and Others v. Commission nationale de l'informatique et des libertés (CNIL), paragraph 62: "With respect to the exception in Article 8(2)(a) of Directive 95/46 and Article 9(2)(a) of Regulation 2016/679, … the consent must be 'specific' and must therefore relate specifically to the processing carried out in connection with the activity of the search engine, and thus to the fact that the processing enables third parties, by means of a search based on the data subject's name, to obtain a list of results including links leading to web pages containing sensitive data relating to him or her."

[1199] WACHTER, S., MITTELSTADT, B. D. & FLORIDI, L., "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation", *International Data Privacy Law,* vol. 7, No. 2, 2017, p. 88.

The fact that the word "predominantly" does not appear in the final version of the text supports the idea that a restrictive interpretation of the phrase "based solely on automated processing" was intended by European authorities.[1200]

This problem has, however, been mostly solved by the guidelines on automated individual decision-making and profiling adopted by the Article 29 Working Party (A29WP).[1201] The A29WP has indicated that human involvement must be meaningful in order to fall beyond the scope of article 22 and it cannot be fabricated.[1202] The A29WP also sets a minimum requirement by which any oversight "should be carried out by someone who has the authority and competence to change the decision" and "should consider all the relevant data".[1203] In any case, it is important to be cautious and not assert that the A29WP has completely solved this issue for there is still room for data controllers to escape the prohibition of article 22 and new issues will undoubtedly arise during the following years.

The reasons behind the regulation contained in article 22 are both dignitary and systemic. A decision based solely on automated processing, including profiling, treats humans as objects, dehumanising them,[1204] consequently, in order to ensure and protect individuals' dignity, human intervention is necessary especially in decision-making processes that produce more relevant effects for them.[1205]

Additionally, article 22 and all the special protections that are built surrounding it are meant to protect individuals from decisions that in theory are more risky in terms of their possible discriminatory or unfair outcomes due to the fact that there is less human control or intervention during data processing.[1206] However, whether human involvement does in fact provide effective safeguards to prevent discrimination is also doubtful for several reasons. On the one hand, humans have been proved to develop "automation bias"[1207], meaning that they tend to believe that the outcomes resulting from automated processes are correct even when

[1200] *Idem*, p. 92.
[1201] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", *cit.*, 2018, p. 21.
[1202] *Ibidem.*
[1203] *Ibid.*
[1204] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1541.
[1205] JONES, M. L., "The right to a human in the loop…", *cit.*, 2017, p. 232.
[1206] BRKAN, M., "Do algorithms rule the world? Algorithmic decision-making in the framework of the GDPR and beyond", *International Journal of Law and Information Technology*, vol. 27, No. 2, 2019, p. 97.
[1207] CITRON, D. K, "Technological due process...", *cit.*, 2008, pp. 1271-1272.

they suspect or are warned of possible system malfunctions.[1208] On the other hand, when algorithms develop biases that lead to discriminatory outcomes they do so through internalising past or existing biases in society that might also be held by the humans who intervene, meaning that, in those cases, human involvement would not actually help to protect individuals' right to equality and non-discrimination.

Thus, even when someone has the authority to change the decision or profile resulting from automated processing, they might overlook or justify elements included in the algorithm that lead to discriminatory decisions. While human intervention does provide an additional safeguard in order to detect errors or biases in the algorithm, the effectiveness of this mechanism will largely depend on the person or team who intervenes in the processing. Furthermore, in some cases in which systemic discrimination leads to accurate statistical discrimination, even if a human has the power to overturn the decision made by the algorithm, if accuracy is preferred over fairness, no changes will be made to the resulting decision.

This does not mean that introducing humans-in-the-loop is not necessary, but that the humans that supervise algorithmic decisions and have the power to overturn them must necessarily understand how the systems they are dealing with work. Additionally, these individuals should also have a significant level of expertise regarding the types of decisions the algorithm is making.[1209]

ii) *Legal or significantly similar effects*

Another of the main problems that are found within the rights and safeguards offered by article 22 is the fact that it refers to decisions which produce "legal effects concerning him or her or similarly significantly affects him or her". The phrase "legal effects" requires the decision to have an effect on an individual's legal rights or status for it to be considered under the scope of article 22.[1210] WACHTER *et al.* interpret this provision restrictively by understanding that, for example, cases related to hiring or credit applications would not be

[1208] PARASURAMAN, R. & MILLER, C. A., "Trust and etiquette in high-criticality automated systems", *cit.*, 2004, p. 52.
[1209] DE-ARTEAGA, M. FLOGLIATO, R. & CHOULDECHOVA, A., "A case for humans-in-the-loop…", *cit.*, 2020, pp. 1-12.
[1210] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", *cit.,* 2018, p. 21; WACHTER, S., MITTELSTADT, B. & FLORIDI, L., "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation", *cit.*, 2017, pp. 92-93.

considered to produce legal effects.[1211] This interpretation is based on the idea that "in most cases the data subject has no legal right to be hired or to be approved for a credit application".[1212]

The A29WP Guidelines offer a few examples in which a decision is considered to produce legal effects. This include cases in which automated decision-making regarding an individual could result in:

- "Cancellation of a contract;
- Entitlement to or denial of a particular social benefit granted by law, such as child or housing benefit;
- Refused admission to a country or denial of citizenship."

WACHTER *et al.*'s interpretation does seem to be therefore consistent with the A29WP Guidelines for, unless a legal instrument recognises individuals' rights in a given situation, the decision made would not be considered to produce legal effects on those affected by it.

The use of the phrase "similarly significantly affects him or her" widens the scope of application of article 22 but generates even greater controversy due to its vagueness. Recital 71 of the GDPR indicates that "automatic refusal of an online credit application or e-recruiting practices without any human intervention" would fall under this category and, consequently the previously cited examples provided by WACHTER *et al.* would in fact be included under the scope of article 22, as their work recognises.[1213]

However, there is still no clear threshold of what is to be considered a similarly significant effect to a legal effect, a difficulty recognised by the A29WP Guidelines[1214] which, nevertheless, do offer a series of elements that could determine a decision based solely on automated processing to fall under the scope of article 22. The Guidelines also indicate that targeted advertising, which has been considered an uncertain case in the literature's

[1211] WACHTER, S., MITTELSTADT, B. & FLORIDI, L., "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation", *cit.*, 2017, p. 93.
[1212] *Ibidem.*
[1213] *Ibid.*
[1214] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", *cit.,* 2018, p. 22.

interpretations of article 22,[1215] may be considered to produce similarly significant effects to legal effects although they warn that a case-by-case approach must be taken.[1216]

2.1.2.5. Analysis of the exceptions to article 22.1 and 22.4

i) *Necessary for entering into, or performance of, a contract*

The first exception to the right not to be subject to decisions based solely on automated processing, including profiling will apply when an individual wants to enter a contract with a data controller for which the latter necessarily requires to resort to automated decision-making either for pre-contractual processing or for processing during the contract's performance. The need to use automated decision-making tools will generally arise from the large amount of data that has to be processed regarding the purpose of the contract. The data subject does not have to provide explicit consent in these cases.

The A29WP Guidelines indicate that the controller must consider the possibility of using a method that is less intrusive and carry out an evaluation to assess whether another such method is available.[1217] The need to weigh whether data collection and the methods used for processing are necessary has also been repeatedly highlighted by CJEU case law.[1218]

However, even though the necessity of resorting to making decisions exclusively based on automated processes should be justified, the exception undeniably provides data controllers with an advantage in structuring contracts that justify fully automated decision-making.[1219]

ii) *Authorised by EU or member state law*

An example of the way in which member states have introduced this exception in their data protection regulation instruments is Section 37 of the German Federal Data Protection Act of 30 June 2017 which establishes a sectorial authorisation for decisions based solely on

[1215] MALGIERI, G. & COMANDÉ, G., "Sensitive-by-distance…", *cit.*, 2017a, p. 247.
[1216] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", *cit.,* 2018, p. 22.
[1217] *Idem*, p. 23.
[1218] CJEU Judgment 16th December 2008, C-524/06, Heinz Huber v. Bundesrepublik Deutschland, Judgment final ruling.
[1219] DREYER, S. & SCHULZ, W., "The General Data Protection Regulation and automated decision-making: will it deliver?", *Bertelsmann Stiftung*, 2019, p. 20. Available on 7th May 2019 at: https://www.bertelsmann-stiftung.de/

automated processing to be made "in the context of providing services pursuant to an insurance contract".

In Spain, Act 40/2015 on the legal regime of the public sector generally authorises the use of automated decision-making by public administrations and establishes that, when automated systems are used in the development of public tasks, the competent body or bodies in charge of defining system specifications, programming, supervision, quality control and, if necessary, auditing of the system and its source code must be predetermined. Additionally, all automated processes must determine the responsible body for the purposes of challenging automated decisions.[1220]

The United Kingdom's case is particularly interesting since it has shifted the general prohibition contained in article 22.1 of the GDPR to a general authorisation establishing general safeguards.[1221]

While not many member states have introduced this exception in their data protection regulation instruments,[1222] they will probably be included in different sectorial regulations throughout the next few years.

iii) *The data subject's explicit consent*

Article 4.11 of the GDPR, states that consent must be "freely given, specific, informed and unambiguous". According to article 7.4 when entering a contract is made conditional to the processing of personal data that is not necessary for said contract this situation shall be taken into special consideration in order to determine whether consent has been freely provided, presuming that consent has not been freely given (Recital 43).

While requiring individuals' consent as a safeguard mechanism for data collection and processing is still necessary, it has proven to be insufficient in order to provide data subjects with enough protections against the abuse of power on the part of data controllers.[1223] Recital

[1220] Article 42.1, Act 40/2015, 1st October 2015, on the legal regime of the public sector.
[1221] Data Protection Act 2018, Section 14.
[1222] MALGIERI, G., "Automated decision-making in the EU member states: The right to explanation and other 'suitable safeguards'", *Computer Law & Security Review*, vol. 35, No. 5, 2019, pp. 23-25.
[1223] EUROPEAN DATA PROTECTION SUPERVISOR, "Towards a new digital ethics: Data, Dignity and Technology", 2015, p. 13: "With all activity potentially always online, the notion of free and informed consent is placed under enormous strain. 'Digital breadcrumbs' are dropped every minute and combined to classify individuals in real

43 specifically addresses situations in which there may be an imbalance of power, considering that this may specially be the case when the controller is a public authority and it is "therefore unlikely that consent was freely given in all the circumstances of that specific situation". Consequently, while public authorities can still rely on data subjects' consent as the lawful basis for data collection and processing, when this exception is the only one that applies to the general prohibition, citizens must be informed that they have no obligation to consent to the collection of their data and those who refuse cannot be denied access to any right or service offered by public institutions.[1224]

The requirements that the consent exception must comply with when it is employed by public authorities sets a double safeguard against discrimination in public service provision. On the one hand, if public authorities must clearly state that individuals will not be deprived access to any service, citizens from lower income backgrounds and who therefore are more in need of certain public services will not feel forced to share their information. On the other hand, the fact that public authorities cannot deny access to any rights or services to individuals who do not provide consent also benefits poorer individuals who generally have greater difficulties in dealing with administrative burdens and barriers to public services.[1225]

Finally, recital 43 indicates that consent is presumed not to be freely given if the controller does not allow separate consent to be given to different personal data processing operations despite it being appropriate in the individual case, or if the performance of a contract, including the provision of a service, is dependent on the consent despite such consent not being necessary for such performance.

iv) *Additional elements that must concur for applying the exceptions to the processing of special categories of personal data*

Article 22.4 indicates that, in order to submit sensitive data to solely automated processing, the controller must comply with paragraphs a or g in article 9.2, that is, have the data

time to create multiple and at times contradictory profiles." These profiles can be circulated in microseconds without individuals' knowledge, and used as the basis for important decisions affecting them.

[1224] ARTICLE 29 WORKING PARTY, "Guidelines on consent under Regulation 2016/679", 17/EN WP259 rev.01, 10th April 2018, p. 6.

[1225] SUNSTEIN, C., "Sludge and ordeals", *Duke Law Review*, vol. 68, No. 8, 2018, p. 1859: "…it is important to focus on the distributional effects of administrative burdens—on whom they are most likely to hurt. As a practical matter, the answer is often the poorest among us".

subject's explicit consent to process her data for one or more specified purposes or processing must take place due to reasons of substantial public interest.

With regard to the first possibility, this requirement offers an additional safeguard with respect to article 22.2.c on two levels. Firstly, due to the fact that member states can establish cases in which the prohibition to process special categories of data cannot be lifted even when consent is provided. Secondly, given that it reinforces the purpose limitation principle. However, the vagueness of the phrase "…or more specified purposes" could lead to the inclusion of general clauses in data collection agreements that would limit the effectiveness of this safeguard.

Processing sensitive categories of data for reasons of substantial public interest is an exception that has been extensively used by member states as it was already included in the DPD as a general exception to the prohibition to process special categories of personal data.[1226] The DPD did, however, not include it as a specific exception (and additional safeguard) to the prohibition of decision-making based solely on the automated processing of special categories of data. In some cases, this exception was transposed into member state legislation in a very extensive way allowing for all processing regulated by an Act to be considered of public interest.[1227] However, within the framework of the GDPR, were this interpretation of article 9.2.g to be valid, this particular additional safeguard set by article 22.4 would be rendered ineffective as the exception of article 22.2.b already allows for solely automated processing to be authorised by Union or member state law.

Consequently, this exception should be interpreted in much more restrictive terms, specially since article 9.2.g indicates that processing "shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject" thereby clearly reflecting the European legislator's objective to particularly restrict decision-making based solely on automated processing, including profiling, of special categories of personal data.

[1226] ARTICLE 29 WORKING PARTY, "Advice paper on special categories of data ('sensitive data')", *cit.*, 20th April 2011, p. 7: "These include for example data processing for the purpose of protecting public security, ensuring the ability of government bodies and authorities to function, preventing serious detriment to the common good, fighting crime, providing equal opportunity as well as social protection and pension systems; for the purpose of scientific research and statistics and for journalistic or artistic purposes. In addition, exceptions cover the areas of preventive medicine, social protection, banking, taxes and customs, labour market policy and unemployment insurance, foreigners and administrative enforcement."
[1227] *Idem*, p. 7.

2.2. PROHIBITIONS IN THE DIRECTIVE FOR DATA PROTECTION IN LAW ENFORCEMENT AND THE CRIMINAL JUSTICE SYSTEM

As Part I conveyed, the use of algorithmic decision-making in law enforcement and the criminal justice system has become increasingly common over the past few years. Automated data processing is used to predict crime through the detection of criminal hotspots[1228] and by identifying traits in individuals' behaviour that are indicators for the commission of illicit activities.[1229] These systems are also used in criminal investigation and prosecution, for instance, in order to determine whether an individual is lying during police interviews[1230] or to detect the veracity of filed reports.[1231] Additionally, tools such as recidivism risk scores, are not only used in prosecution strategies[1232] but also to make decisions regarding the execution of criminal penalties.[1233]

The introduction of these techniques by law enforcement and criminal justice institutions results from the amount of data that these public organisations collect and hold and the opportunities that the development of information processing technologies offer in improving the efficiency that is increasingly required from public administrations.[1234]

There are certain tensions that arise between the protection of individuals' personal data and the public interests that must be protected by the state.[1235] These conflicts are especially tangible with regard to the police and criminal justice systems seeing as the tasks that must be carried out to guarantee law enforcement unavoidably interfere in the free development of human rights, including the right to privacy and personal data protection.[1236] For this reason, the EU considered appropriate to dedicate a separate regulatory instrument that was

[1228] O'NEIL, C., *Weapons of Math Destruction…*, *cit.*, 2017, p. 85.
[1229] FERGUSON, A. G., "Big data and predictive reasonable suspicion", *cit.*, 2015, p. 335; CAPDEFERRO VILLAGRASA, O., "El análisis de riesgos como mecanismo central de un sistema efectivo de prevención de la corrupción…", *cit.*, 2018, pp. 1-7.
[1230] IBORDERCRTL, "Technical Framework", 2016. Available on 16th May 2019 at: https://www.iborderctrl.eu/
[1231] QUIJANO-SÁNCHEZ, L. *et al.*, "Applying automatic text-based detection of deceptive language to police reports…", *cit.*, 2018, pp. 155-168; KOLOTÚSHKINA, N., "VERIPOL…", *cit.*, 2018.
[1232] FERGUSON, A. G., "Predictive prosecution", *Wake Forest Law Review*, vol. 51, No. 3, 2016, pp. 705-744.
[1233] KEHL, D., GUO, P. & KESSLER, S., "Algorithms in the Criminal Justice System…", *cit.*, 2017, p. 2.
[1234] OSWALD, M. *et al.*, "Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality", *Information & Communications Technology Law*, vol. 27, No. 2, 2018, p. 223.
[1235] BRKAN, M., "Do algorithms rule the world?...", *cit.*, 2017, pp. 23-24.
[1236] CARUANA, M. M., "The reform of the EU data protection framework in the context of the police and criminal justice sector: harmonisation, scope, oversight and enforcement", *International Review of Law, Computers & Technology*, 2017, p. 1.

specifically adapted to the issues regarding the processing of personal data in law enforcement and the criminal justice system.

2.2.1. Harmonisation and scope of application

Pursuing the harmonisation through a Directive is much less effective than doing so through a Regulation.[1237] However and although it is true that the GDPR offers a more complete regulatory framework than Directive 2016/680 for data protection in law enforcement, the legislative technique employed in the former provides member states with a significant scope of choice when applying the GDPR in their own jurisdictions.[1238]

Another element that may bring issues regarding the consistent application of Directive 2016/680 for data protection in law enforcement, is the fact that the differences in the scope of application between said instrument and the GDPR are not completely delimited. The Directive applies to "competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security" (articles 1 and 2) where 'competent authority' means (article 3.7):

> "(a) Any public authority competent for the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security; or
>
> (b) Any other body or entity entrusted by member state law to exercise public authority and public powers for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security."

The application of paragraph (b) within member states will probably result in divergent personal scopes of application when the Directive is transposed, leading to the application of the GDPR to certain types of data processing in a member state while the same types of data processing will be covered by the national laws passed in application of Directive 2016/680 for data protection in law enforcement in other member states.[1239] Additionally, where

[1237] *Idem*, p. 4.
[1238] MALGIERI, G., "Automated decision-making in the EU member states…", *cit.*, 2019.
[1239] CARUANA, M. M., "The reform of the EU data protection framework…", 2017, p. 5.

national security issues or criminal procedure rules apply, member states are not directly subjected to the Directive and, although this does not mean that it is inapplicable, the extent to which member states will rely on their internal rules will result in further fragmentation in the application of the Directive.[1240]

2.2.2. Processing special categories of personal data within the scope of Directive 2016/680 for data protection in law enforcement

Article 10 of the Directive, like article 9 of the GDPR, regulates the processing of special categories of personal data. The wording in article 10 of the Directive for data protection in law enforcement and the criminal justice system differs from the GDPR as it does not explicitly use the term "prohibit" but indicates that processing of special categories of personal data is only allowed "where strictly necessary, subject to appropriate safeguards for the rights and freedoms of the data subject, and only (a) where authorised by Union or member state law; (b) to protect the vital interests of the data subject or of another natural person; or (c) where such processing relates to data which are manifestly made public by the data subject."

More exceptions to the general prohibition are listed in the GDPR than in the Directive. This is explained by the greater risks that data processing and automated decision-making in law enforcement pose to individuals' fundamental rights, as well as the difference in the type of legal instrument employed, since Directives set less specific provisions than Regulations in order for them to be adapted to each member state's legal system.

While Article 10 effectively operates as a prohibition, the exception included under paragraph (a) provides member states which a much wider room for manoeuvre when introducing exceptions by which the authorities whose activities fall under the application of Directive 2016/680 for data protection in law enforcement can process special categories of data.

With respect to choosing an extensive or restrictive interpretation of the special categories of data, it is important to highlight that the extensive interpretation of the special categories of data contained in Article 10 of Directive 2016/680 for data protection in law enforcement would probably lead to the impossibility of using recidivism risk prediction mechanisms

[1240] *Idem*, pp. 8-9.

given the fact that much of the information used in them correlates with race or ethnicity unless, as it will probably be the case, the exception regarding member state explicit authorisation is employed.

2.2.3. The prohibition of decisions based solely on automated processing, including profiling

With regard to article 11 of the Directive, which is equivalent to article 22 of the GDPR, it is important to highlight that unlike the latter, article 11 is explicitly structured as a general prohibition and not as a right of the data subject. Said prohibition covers decisions "based solely on automated processing, including profiling, which produces an adverse legal effect concerning the data subject or significantly affects him or her." In general, data processing tools used in law enforcement and the criminal justice system are still used as part of processes in which humans intervene.[1241] Consequently, the application of article 11 will be quite limited for the time being although it will acquire an increasing relevance as machine learning algorithms occupy a larger part of processes in law enforcement and the criminal justice system.

The GDPR does not require adverse legal effects in order to include automated processing under the scope of the prohibition set in article 22. However, the fact that most decisions made by law enforcement bodies and the criminal justice system produce effects on the sphere of fundamental rights means that there will almost always be a risk of adverse legal effects taking place. Even when automated processing is used in order to, for example, include individuals in reinsertion programmes, which is a positive effect, non-inclusion in said programmes will cause harms to individuals.

Exceptions to the general prohibitions can only be introduced by member state or EU legislation, which is coherent with the fact that the regulatory instrument is a Directive and that member states have autonomy in structuring and regulating their law enforcement and criminal justice systems. Moreover, the principle set by articles 4.1(a) and 8 demands that all data processing must be lawful. Thus, in line with this principle the only cases in which decisions based solely on automated processing, including profiling, can be carried out must necessarily have legal footing.

[1241] FERGUSON, A. G., "Predictive prosecution", *cit.*, 2016, pp. 705-706.

The legal instrument that allows decision-making based solely on automated processing, including profiling, must also necessarily establish appropriate safeguards. While this requirement is very similar to the one set by article 22 of the GDPR, the only minimum safeguard explicitly contained in article 11 of the Directive is the right to obtain human intervention. In this context, it is necessary to consider Recital 38 of the Directive which indicates that decision-making based solely on automated processing, including profiling, "should be subject to suitable safeguards, including the provision of specific information to the data subject and the right to obtain human intervention, in particular to express his or her point of view, to obtain an explanation of the decision reached after such assessment or to challenge the decision". Consequently, within the regulation contained in the Directive the data subject's rights to express her point of view, obtain and explanation and challenge the decision are all necessary expressions of the right to obtain human intervention. Seeing as these are forms of due process rights granted to the data subject they will be analysed in more detail in the following section.

With regard to the issue of discrimination that is central to this research, the second paragraph of article 11 indicates that automated processing of special categories of data is prohibited unless "suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place". The only difference between this prohibition and the one in paragraph 1 of the article is that the phrase "*legitimate interests*" is included here. Consequently, this provision serves more as a reminder of the fact that more risks are generated for individuals when special categories of data are processed and that authorities must be particularly careful when processing said types of data than as actual protection against said processing.

As Part I illustrated, some of the machine learning algorithms used in law enforcement and the criminal justice system have been proven to discriminate against members of disadvantaged groups and reinforce existing patterns of systemic discrimination which have especially serious consequences given the fact that fundamental rights and freedoms can be heavily restricted in the areas covered by Directive 2016/680 for data protection in law enforcement. Concerns with regard to the discriminatory outcomes that may result from the use of newly developed data processing technologies under the scope of the Directive are reflected in article 11.3 as it states: "profiling that results in discrimination against natural

persons on the basis of special categories of personal data […] shall be prohibited, in accordance with Union law".

2.3. SHORTCOMINGS IN THE PROHIBITIONS CONTAINED IN THE EU DATA PROTECTION FRAMEWORK

This section conveys the fact that traditional privacy regulations are not sufficiently adapted to the new issues and risks that arise from the computing capabilities of data processing technologies. Passing regulatory instruments such as the GDPR or Directive 2016/680 for data protection in law enforcement is necessary in order to properly deal with machine learning algorithms and their increasing pervasiveness. However, privacy protection through prohibitions as it is shaped in the aforementioned European legislation is far from providing individuals with enough safeguards against the risks of discrimination that algorithmic decision-making poses.

While more general shortcomings in EU regulatory instruments will be approached later on, it is important to highlight that [1242] the GDPR and the Directive for data protection in law enforcement and the criminal justice system fall short of offering effective protections to individuals in order to prevent discriminatory practices through anticlassification tools.

The fact that processing of special categories of data is prohibited and this prohibition is especially reinforced when said processing is solely automated seem to depict the existence of an ironclad system that has been put in place in order to protect members of traditionally oppressed groups from the risks resulting from data processing technologies. However, said prohibitions will be inapplicable in many cases if the exceptions to them are extensively interpreted and applied by member states. Additionally, and more importantly, the fact that special categories can be inferred from other pieces of data makes it very difficult to draw the line regarding the extent to which the prohibitions should apply.

Finally, the EU framework considers human intervention to offer certain safeguards to individual rights in automated decision-making that will render the prohibitions in article 22 inapplicable. However, for human intervention to be effective, it is necessary to ensure that the individuals supervising automated tools understand them and the context in which decisions are being made.

[1242] BALKIN, J. M. & SIEGEL, R. B., "The American civil rights tradition...", *cit.*, 2003, p. 10.

3. TECHNOLOGICAL DUE PROCESS[1243] RIGHTS

Amongst the many issues that come up when analysing the use of data processing technologies such as machine learning algorithms is the fact that individuals are not actually aware of how much of their data they are sharing and the potential that these technologies actually have.[1244] This is one of the reasons why part of the scholarship suggest framing and regulating individual rights for protection against algorithmic discrimination and other unfair treatments resulting from automated decision-making through the informational self-determination paradigm. The logic behind these proposals is to empower individuals by providing them with sufficient information and knowledge regarding data processing technologies so that they are able to choose when and what data to share.

One of the proposals that have been put forward in order to solve the aforementioned problems from the perspective of empowering individuals is presented by MALGIERI and COMANDÉ, who suggest approaching data through the paradigm of "data management",[1245] to prevent and solve situations in which the automated processing of an individual's health (or health related) data may discriminate members of especially vulnerable groups.

In a similar sense, ILLINGWORTH, suggests that these approaches should especially focus on individuals' power over the use given to their data without removing controls on data collection and storage seeing as the current possibilities offered by new big data technologies cannot be properly managed by traditional approaches to the handling of personal information.[1246] The objective that underlies these proposals is to empower individuals[1247] by increasing their awareness through transparency[1248] and providing them with mechanisms to chose how, when and what information to share and for what purposes.[1249]

This informational self-determination approach is actually the way in which data protection rules have traditionally been developed, both at the European Union and member state level. For example, the Spanish Constitutional Court in its 292/2000 Judgment stated the following:

[1243] CITRON, D. K., "Technological due process", *cit.*, 2008.
[1244] MALGIERI, G. & COMANDÉ, G., "Sensitive-by-distance…", *cit.*, 2017a, p. 230.
[1245] *Idem*, p. 231.
[1246] ILLINGWORTH, A. J., "Big data in I-O psychology…", *cit.*, 2015, pp. 569-570.
[1247] In fact, private initiatives in order to help develop awareness on this issue are already being developed. See MYDATA, "Who we are", 2018. Available on 5th May 2019 at: https://mydata.org/about/
[1248] Transparency, the right to explanation and legibility will be approached in the Chapter dedicated to transparency.
[1249] HILDEBRANDT, M. & KOOPS, B. J., "The challenges of ambient law and legal protection in the profiling era", *cit.*, 2010, p. 429.

> "The fundamental right to data protection seeks to guarantee the individual a power of control over their personal data, their use and purpose, with the objective of preventing data trafficking that is illicit and harmful for the dignity and rights of the individual."

Much of the protection framed from the individual rights perspective in the EU, and specifically in the GDPR, consequently aims towards providing data subjects with information regarding how their data is being or has been collected and used as well as mechanisms to proactively manage their data thus reflecting the importance of the individual informational self-determination approach in the current framework of protection from data processing from the perspective of individual rights. Hence, the European Data Protection framework grants individuals informational passive rights that correlatively entail obligations for data processors and controllers to inform and provide some manner of explanation to data subjects with regard to processing and automated or semi-automated decision-making.

This approach goes in line with the creation of a "technological due process"[1250] that has been defended by part of the scholarship as a necessary mechanism in order to properly defend individuals from the risks that automated decision-making and profiling pose on their privacy,[1251] their right to not be discriminated[1252] and fairness in general.[1253] This literature, which mostly originates in the US, has largely focused on the deficiencies of the US sectorial privacy approach to deal with the issues and risks produced by the new capabilities of data collection and processing tools.[1254] Additionally, it is argued that existing sectorial privacy regulations in the US, which in many cases place their focus on data collection, processing and disclosure prohibitions, can be easily bypassed by the computational power of data processing technologies.[1255]

Proponents of technological due process thus argue that shifting the way in which traditional privacy regulations are structured is necessary in order to provide individuals with a comprehensive set of rights that allow them to react and protect themselves from the collection and processing of their data.[1256] In fact, proponents of technological due process

[1250] CITRON, D. K., "Technological due process", *cit.*, 2008; CITRON, D. K. & PASQUALE, F., "The scored society…", *cit.*, 2014; CRAWFORD, K. & SCHULTZ, J., "Big data and due process: toward a framework to redress predictive privacy harms", *Boston College Law Review*, vol. 55, No. 1, 2014, pp. 93-128.
[1251] CRAWFORD, K. & SCHULTZ, J., "Big data and due process…", *cit.*, 2014, p. 95.
[1252] CITRON, D. K. & PASQUALE, F., "The scored society…", *cit.*, 2014, pp. 14-15.
[1253] *Idem*, p. 19.
[1254] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1554.
[1255] CRAWFORD, K. & SCHULTZ, J., ., "Big data and due process…", *cit.*, 2014, p. 106.
[1256] *Idem*, pp. 109-110.

rights defend that protecting privacy through technological due process rights can also help to prevent and deal with instances of discrimination (and other harms) resulting from data processing. By requiring organisations to offer prior notice of the data that will be processed and *ex post* information to individuals of the pieces of data that have been used in making a decision about them, it will be easier to control and ensure that fundamental rights, including the right to equality and non-discrimination, are protected.[1257]

This is precisely the basis upon which the GDPR has been built seeing as, it does not just offer general prohibitions that operate in order to protect individuals but also a comprehensive catalogue of rights and correlative obligations for data controllers and processors that provide data subjects with the possibility of obtaining information and explanations regarding the collection and processing of their data, of intervening and being heard during data collection and processing and of challenging decisions. Consequently, the GDPR includes many of the elements that are considered necessary for a technological due process.[1258]

This section addresses the way in which the GDPR, Directive 2016/680 for data protection in law enforcement and US regulatory instruments have included due process rights as part of their framework, analysing the way in which the regulatory authorities in the EU and the US have decided to weigh the conflicting interests between transparency and public or private interests from the perspective of individuals' rights to an algorithmic due process. Once the different regulatory solutions and proposals regarding individual transparency have been put forward, the section moves on to discuss the possibilities that individuals have to actively intervene and challenge the way in which their data is processed by automated systems.

3.1. TRANSPARENCY: THE RIGHTS TO INFORMATION, ACCESS AND EXPLANATION

Transparency-related rights are the cornerstone to the catalogue of rights contained in the EU data protection framework. The level of transparency which rights to information, access and explanation grant individuals will have a huge influence in the exercise of the rights to be heard and challenge decisions and their specific materialisations. This is highly important given the fact that the degree of transparency and thus, the extent to which individuals will be

[1257] CRAWFORD, K. & SCHULTZ, J., "Big data and due process…", *cit.*, 2014, p. 111; ZUIDERVEEN BORGESIUS, F., "Discrimination, artificial intelligence and algorithmic decision-making", Strasbourg, Directorate General of Democracy, Council of Europe, 2018, p. 21.
[1258] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1586-1590.

able to exercise their other rights, will depend on the weight placed on the principle of transparency as opposed to other conflicting interests such as intellectual property rights or national security.

Due process approaches argue for the recognition of information (or notice) rights regarding collection and processing of data as well as the purposes for which these activities are being carried out.[1259] Additionally, it is argued that these rights should also offer individuals the possibility of accessing their information in order to ensure that the collected and processed data is accurate.[1260] Information rights would be completed with a right to explanation regarding input variables employed in the processing and the methodology used,[1261] and detailed information regarding all of the decisions made in a certain procedure.[1262]

Part of the literature has also claimed offering individuals the possibility of experimenting with the models used to make decisions that involve them as an aspect of the rights to information, access and explanation.[1263] For example, CITRON and PASQUALE argue that if an individual has access to an interface with all of her credit information in which she can modify different elements in order to see what changes to her credit score will result from, for example, choosing to pay one bill or another, she will be able to make informed decisions with regard to her financial behaviour.[1264]

Calls for systemic transparency and accountability have also been made from this perspective in order to ensure that rights could be properly enforced.[1265] Since said elements of the technological due process respond to systemic regulation instead of individual right protection they will be approached in the following section.

[1259] CRAWFORD, K. & SCHULTZ, J., "Big data and due process…", *cit.*, 2014, pp. 125.
[1260] *Ibidem*.
[1261] *Idem*, p. 126.
[1262] CITRON, D. K., "Technological due process", *cit.*, 2008, p. 1305; CITRON, D. K. & PASQUALE, F., "The scored society...", *cit.*, 2014, p. 28.
[1263] CITRON, D. K. & PASQUALE, F., "The scored society...", *cit.*, 2014, pp. 28-30; HILDEBRANDT, M., "The dawn of a critical transparency right for the profiling era", in BUS, J. *et al.* (eds.), *Digital Enlightenment Yearbook 2012*, Amsterdam, IOS Press, 2012, p. 53.
[1264] CITRON, D. K. & PASQUALE, F., "The scored society...", *cit.*, 2014, p. 29.
[1265] CITRON, D. K., "Technological due process", *cit.*, 2008, pp. 1310-1312; CITRON, D. K. & PASQUALE, F., "The scored society...", *cit.*, 2014, pp. 20, 24-28, 33.

3.1.1. Information, access and explanation rights in the GDPR

The transparency principle that must be present at all stages of data processing is contained in article 5 of the GDPR.[1266] The GDPR sets a series of transparency related obligations for data controllers and recognises rights for data subjects to ensure that the latter have enough information regarding the processing of their data before a decision or profile has been carried out. Article 12 provides the general rules that data controllers should follow, establishing a general obligation to carry out the necessary measures to ensure that all rights to information and access recognised under the GDPR can be made effective. The rights to information and access explicitly recognised in articles 13 to 15 are the specific materialisations of the broader rule of transparency that directly protect data subjects.

Additionally, the way in which consent is structured under the GDPR also provides a source of transparency. Consent is sometimes required in order to carry out data processing and acts an as exception to the general prohibitions contained in articles 9 and 22 of the GDPR. In order to consider that consent has been freely obtained, the individual must be correctly informed and "understand exactly what they are consenting to",[1267] meaning that the controller must offer her at least some degree of explanation regarding the way in which her data will be processed.[1268]

Transparency and the individual rights that derive from it are vital in order to help reduce the informational asymmetries that exist between data controllers and data subjects,[1269] especially when automated tools are used.[1270] These asymmetries result from the general opacity and complexity of machine learning algorithms; from the large amounts of data that individuals inadvertently share and are collected and used by actors in the private and public sector and, finally, from the general positions of power that data processors and controllers hold within society and which provide them with institutional and legal mechanisms, such as intellectual

[1266] As it was previously mentioned, the systemic implications of this general principle will be approached in Part III. This Part, however, focuses on the individual rights of data subjects and correlative obligations of controllers and processors.
[1267] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", *cit.,* 2018, p. 23.
[1268] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1590.
[1269] MALGIERI, G. & COMANDÉ, G., "Sensitive-by-distance…", *cit.*, 2017a, p. 230.
[1270] BAYAMLIOĞLU, E., "Transparency of automated decisions in the GDPR…", *cit.*, 2018, p. 17.

property rights and trade secrets or public interests in the case of the public sector, that help them preserve the opacity of the tools and processes they use.[1271]

The other objective that transparency aims to accomplish is providing individuals with enough resources to challenge decisions or profiles resulting from the processing of an individual's data.[1272] A proper explanation of the underlying logic leading to the decision or profile made is necessary in order for a data subject to properly defend her position during a review process.[1273] However, an obvious limitation in cases of discrimination is that the data subject only has access to her own data but not that of other subjects, which may heavily hinder the possibility of individuals to become aware of the discrimination that they are being subjected to, unless controllers are forced to provide some form of aggregate information and the individual's relative position in relation to the other subjects whose data is analysed.[1274]

3.1.1.1. The right to be informed

i) *The intended purposes of the processing*

A right that is recognised to all data subjects, whether they are subject to decisions based solely on automated processing, including profiling, or not, is the right to be informed of the purpose for which their data will be used (articles 13.1.c and 14.1.c). The recognition of this right is consistent with the principle of purpose limitation that must rule all data processing according to article 5 of the GDPR.[1275]

By establishing data subjects' right to be informed of the purposes for which their data will or is being processed individuals are provided with a key piece of information regarding whether the controller has a legal basis for processing her data that could be used in case they have to contest the decision or profile resulting from their processing of their data.[1276] Moreover, an additional safeguard is placed in order to provide some constraints to the uncontrolled spread of personal data processing.

[1271] *Ibidem.*
[1272] ZERILLI, J. *et al.*, "Transparency in algorithmic and human decision-making: is there a double standard?", *Philosophy and Technology*, 2018, p. 2.
[1273] *Ibidem.*
[1274] HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, pp. 1173-1174.
[1275] BAYAMLIOĞLU, E., "Transparency of automated decisions in the GDPR…", *cit.*, 2018, p. 25.
[1276] *Ibidem.*

The general prohibition included in article 22.1 and 22.4[1277] and the additional safeguards provided by the GDPR to decisions based solely on automated decision-making, including processing, have a very limited scope of application that can be easily circumvented by introducing a "human in the loop" that does not necessarily provide additional guarantees to individuals' rights. Consequently, recognising the right to be informed of the intended purpose for which the data will be processed is essential in order to guarantee at least some degree of protection for situations not covered under article 22 and to which the rights to information and access (and possibly of explanation) that will be discussed in the following pages do therefore not apply.

ii) *Meaningful information about the logic involved, significance and envisaged consequences*

Both articles 13.2.f and 14.2.f state that the controller must provide the data subject information regarding "the existence of automated decision-making, including profiling, referred to in article 22.1 and 4 and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject".

The main difference between articles 13 and 14 is that article 13 applies when the personal data are collected from the subject, whereas articles 14 applies when the data have not been obtained directly form the data subject. The rights recognised in said articles are reinforced in

[1277] Article 22 GDPR: "1.The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her. 2.Paragraph 1 shall not apply if the decision: (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (c) is based on the data subject's explicit consent. 3.In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision. 4.Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place."
Article 9.1 and 2.a and g GDPR: "1.Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited. 2.Paragraph 1 shall not apply if one of the following applies: (a) the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject; (g) processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject".

the case of profiling given the fact that according to the A29WP Guidelines, data subjects must be informed about profiling "regardless of whether solely automated individual decision-making based on profiling takes place".[1278] The fact that data subjects have more information rights with regard to profiling than with respect to other types of data processing that are not automated is highly relevant given that profiling is generally used as the basis for automated decision-making or as a complementary tool for it.

It is important to acknowledge that these articles initially place focus on informing individuals that they are being subjected to automated decision-making or profiling and only then on the logic involved in said processing[1279]. This approach makes sense since the Regulation aims to protect individuals through offering them the possibility of halting the automated decision-making or profiling if there is not enough legal base for it or of requiring human intervention. If neither of this options is adopted the data subject can still contest elements involved in the decision-making process or the resulting decision.

In order for the data subject to receive "meaningful information about the logic involved", she must be provided enough information to understand the way in which the automated decision-making process works and, more specifically, the rationale and criteria behind the process.[1280] In doing so, the controller must inform the data subject on the main elements of data used in the decision, their source and their relevance.[1281]

Additionally, the information provided with regard to the significance and envisaged consequences of the automated decision-making process or profiling should be delivered with examples that illustrate the possible outcomes of said processing in order for the data subject to actually understand the extent to which the automated process or profiling could affect her the information.[1282]

Finally, the fact that article 12 of the GDPR, which establishes the general information obligations for controllers indicates that the information has to be provided "in a concise, transparent, intelligible and easily accessible form, using clear and plain language", comes to

[1278] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", *cit.*, 2018, p.17.
[1279] DREYER, S. & SCHULZ, W., "The General Data Protection Regulation and automated decision-making…", *cit.*, 2019, p. 23.
[1280] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", *cit.*, 2018, p. 25.
[1281] *Idem*, p. 26.
[1282] *Ibidem.*

support the notion that information must be delivered in a way that helps the data subject to fully understand the pieces of personal and non-personal information involved in the process, the ways in which they are used and the consequences this may bring for her.

The debate on the information that would have to be disclosed under articles 13 and 14 was mainly solved after A29WP issued the guidelines on automated decision-making, which clearly stated that the data subject had to understand how the decision had been made and the basis for it.[1283]

3.1.1.2. The right to access

Article 15.1 of the GDPR regulates the right of the data subject to request confirmation that her personal data are being processed. If the answer is positive she can access her personal data and, if she is being subjected to a solely automated decision-making process, including profiling, "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject". This right is therefore structured in a very similar way to the right to information recognised under articles 13.2.f and 14.2.f.

The right to access thus provides the individual a reinforcement mechanism to the passive right to information that controllers must comply with according to articles 13 and 14. The data subject can exercise her right at any moment throughout the processing of information thereby enhancing the possibilities she has to collect all the information possible that may result useful were she to contest the resulting decision or profile.

The right to access can be particularly relevant with regard to the resources that the data subject will have when and if she wishes to contest the decision or profile seeing as the A29WP guidelines state that, amongst the information provided when this right is exercised, the controller should provide information on "factors taken into account for the decision-making process, and on their respective 'weight' on an aggregate level".[1284] Information on the way different elements are considered by the algorithm will be useful in order to provide a solid argument that challenges the decision for it will enable the data subject to properly attack the rationale behind the algorithms decision.

[1283] *Idem*, p. 27.
[1284] *Ibidem*.

3.1.1.3. The right to explanation

A right to explanation is not explicitly recognised under the GDPR. Furthermore, part of the scholarship still argues that this right does not exist under the GDPR and that the rights to information and access only offer a general explanation about the automated processing system but not a specific explanation of the underlying logic to the decision made regarding the individual.[1285]

The right to obtain an explanation when decisions are based solely on automated processing, including profiling, is only explicitly recognised in Recital 71 of the GDPR, which does not have a binding nature.[1286] However, a systematic interpretation of articles 12 to 15 along with Recital 71 and the A29WP Guidelines, all of which have been analysed in the previous sections, does seem lead to the conclusion that a right to explanation does exist in the GDPR.[1287] The exact information that should be disclosed under this right is however still not clear[1288] but it should, for example, include "the categories of data used to construct a profile"[1289] and be sufficient for the data subject to understand the reasons for the decision or profile,[1290] meaning that a general explanation of the way the automated decision-making or profiling tool works is not enough to consider that the transparency rule has been complied with.

In any case, the right to explanation does not exist with regard to cases that do not fall under the scope of article 22 of the GDPR. As it was previously stated, circumventing the application of article 22 could be a very easy task under certain circumstances. In cases in which a human in the loop that can overturn algorithmic decisions is introduced, the full right to explanation, including the logic involved in the decision or profile will not exist.[1291]

Additionally, even when the right to explanation applies, there are some internal and external limits that may render it partly ineffective in some cases.

[1285] WACHTER, S., MITTELSTADT, B. D. & RUSSELL, C., "Counterfactual explanations without opening the black box…", *cit.*, 2018, pp. 868-869.
[1286] CJEU Judgment 13th July 1989, C-215/88, Casa Fleischhandel v. Bundesanstalt für landwirtschaftliche marktordnung.
[1287] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1593.
[1288] *Ibidem.*
[1289] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", *cit.*, 2018, p. 17.
[1290] *Idem*, p. 25.
[1291] *Idem*, pp. 17, 31.

i) *Internal limits to the right to explanation*

The difficulty of achieving something similar to transparency for automated decision-making processes is twofold. On the one hand, the complex and multi-level design of machine learning algorithms makes it so much harder to understand their underlying logic.[1292] On the other hand, the fact that self-learning models can be fed and process massive datasets means that many of the variables used by these algorithms cannot be translated for human beings to understand.[1293]

The complexity of the processes used cannot serve as an excuse for controllers to fail to provide adequate information[1294] in order to ensure the effectiveness of the data subjects' rights recognised under Recital 71 and articles 12 to 15. However, in some cases these rights and obligations may be impracticable, especially when dealing with machine learning algorithms, seeing as the ways in which the algorithms will weigh in the different variables and draw correlations is unexpected.[1295] Moreover, in many cases in which self-learning tools are used the actual logic underlying the resulting decision or profile is not even understood by the creators of the algorithm.[1296] Consequently, data subjects will not always have the possibility of obtaining information on the underlying logic to the decisions made by automated systems.

ii) *External limits to the right to explanation*

a. The conflict with trade secrets and intellectual property

Intellectual property law, including trade secrets, aims to foster creativity and productivity in order to maximise social welfare by protecting the innovations produced by individuals or firms.[1297] By ensuring individuals and businesses that their creations will be correctly

[1292] GOODMAN, B. W., "A step towards accountable algorithms?...", *cit.*, 2016, p. 4.
[1293] *Ibidem.*
[1294] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", *cit.*, 2018, p. 25.
[1295] ZERILLI, J. *et al.*, "Transparency in algorithmic and human decision-making…", *cit.*, 2018, pp. 3-4.
[1296] BAYAMLIOĞLU, E., "Transparency of automated decisions in the GDPR…", *cit.*, 2018, p. 17.
[1297] WORLD INTELLECTUAL PROPERTY ORGANISATION, "What is intellectual property?", 2004, p. 3. Available on May 22nd 2019 at: https://www.wipo.int/publications/es/details.jsp?id=99&plang=EN

protected and that they will reap most of the benefits that result from them, they are incentivised to invest more resources in order to continue with innovative processes.[1298]

Within the scope of data processing, intellectual property rights enter into conflict with individuals' rights given that solely or partly automated profiling and decision-making can hamper the latter. In order to ensure that individuals can challenge profiles and automated decisions that affect them, they must be provided with information that explains why the data processing tools used yielded the results that are to be contested.

One of the main limits to algorithm transparency is established in Recital 63 of the GDPR which indicates that the right of data subjects to access their data being used by the controller "should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software". The Recital prevents this possibility from becoming a loophole through which data controllers may avoid complying with transparency requirements by indicating that "the result of those considerations should not be a refusal to provide all information to the data subject". The A29WP Guidelines specially emphasise the fact that controllers and processors cannot rely on intellectual property rights, including trade secrets, in order to deny individuals their transparency related rights.[1299]

BRKAN makes a very strong case regarding the fact that intellectual property rights cannot actually undermine the right to explanation or algorithmic transparency in general, as it is constructed in the GDPR.[1300] She argues that it is very unlikely for an algorithm to be patented and even if it were the patent holder would have to "disclose the composition and the modalities of functioning of an algorithm".[1301] Similarly, copyright protections would not serve as a mechanism for data processors o controllers to bypass transparency related obligations given that it is not clear that algorithms can be protected through copyright and,

[1298] STOPFAKES.GOV (INTELLECTUAL PROPERTY RIGHTS INFORMATION & ASSISTANCE), "Why is intellectual property important?", July 7th 2016. Available on May 22nd 2019 at: https://www.stopfakes.gov/article?id=Why-is-Intellectual-Property-Important: "Intellectual property protection is critical to fostering innovation. Without protection of ideas, businesses and individuals would not reap the full benefits of their inventions and would focus less on research and development. Similarly, artists would not be fully compensated for their creations and cultural vitality would suffer as a result."
[1299] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", *cit.*, 2018, p. 17.
[1300] BRKAN, M., "Do algorithms rule the world?...", *cit*, 2017, pp. 21-23.
[1301] *Idem*, p. 22.

in any case, the EU Computers Directive also establishes certain transparency rights for programme users.[1302]

Finally, article 5.d of the EU Directive on Trade Secrets,[1303] establishes that trade secrets can be disclosed "for the purpose of protecting a legitimate interest recognised by Union or national law".[1304] Seeing as the rights derived from the principle of transparency are comprehended under the GDPR and thus generally aimed towards the protection of individuals against the processing of their data, which is a fundamental right recognised in articles 8.1 of the Charter of Fundamental Rights of the European Union and 16.1 of the Treaty on the Functioning of the European Union (TFEU), transparency of data processing is a legitimate interest that should generally outweigh trade secret protections.

Additionally, even if trade secret protections were applicable in some cases, which is precisely what must be drawn from Recital 63 of the GDPR, in order to comply with the transparency principle and the rights and obligations that derive from it, DIAKOPOULOS argues it is not necessary for firms to reveal the source code or expose their systems since providing a few pieces of key information would be enough to ensure algorithmic accountability.[1305]

However, the role that trade secrets might play in limiting the right to explanation as well as the other rights derived from the principle of transparency in the GDPR should not be downplayed. For example, German courts have ruled with regard to credit scoring that data subjects' right to an explanation of the logic involved was limited by trade secret protections.[1306] In this sense, it is important to highlight that the degree of specificity achieved by the European legislator when balancing transparency and IP rights is far from being ideal. This lack of specificity can serve as an advantage for controllers and processors that can either adopt a very restrictive interpretation of transparency-related rights thereby not

[1302] Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs. Recital 16: "A person having a right to use a computer program should not be prevented from performing acts necessary to observe, study or test the functioning of the program, provided that those acts do not infringe the copyright in the program".
[1303] Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure.
[1304] BRKAN, M., "Do algorithms rule the world?...", *cit*, 2017, p. 23.
[1305] DIAKOPOULOS, N., "Accountability in algorithmic decision making", *Communications of the ACM*, 2016, vol. 59, No. 2, 2016, pp. 58-59: "Complete source-code transparency of algorithms, however, is overkill in many if not most cases. Instead, the disclosure of certain key pieces of information, including aggregate results and benchmarks, would be far more effective in communicating algorithmic performance to the public".
[1306] WACHTER, S., MITTELSTADT, B. D. & FLORIDI, L., "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation", *cit.*, 2017, p. 87.

providing individuals with actually useful information or can lead to releasing such large amounts of information with questionable legibility that it is impossible to discern what are the relevant elements in order for an individual to challenge a decision.[1307]

b. State secrets and public interests

The public sector's mandate for accountability is far stronger than the transparency related obligations that the private sector has.[1308] Moreover, the discussion on whether the right to an explanation exists is less problematic when applied to public sector decisions due to "the obligation of the administration to give reasons for its decisions".[1309] However, in some cases, exceptions to transparency mandates enable accountability will necessarily be made in order to safeguard certain public interests.[1310] This issue will be approached in more detail when transparency in the law enforcement directive is addressed.

iii)*How the right to explanation can be made effective*

MALGIERI and COMANDÉ consider that what the GDPR really recognises is a right to legibility that allows individuals to understand how the algorithms that can influence their lives work and the extent of the impact they could have[1311] and develop a set of questions regarding automated decision-making the answers to which controllers should provide data subjects in order to comply with the transparency-related rights contained in the GDPR.[1312] However, they do not solve the problems that may arise from conflicts with trade secrets.

BINNS *et al.*[1313] offer a series of possibilities for controllers to comply with transparency obligations in order to ensure the effectiveness of data subjects' right to explanation: 1) Disclosing information on the variables used as well as their positive or negative weight regarding the final decision; 2) informing individuals how much a variable would have to change in order for them to be included in a different output class; 3) offering an explanation of the algorithm based on an example of training data that closely resembles the data

[1307] ANNANY, M. & CRAWFORD, K., "Seeing without knowing…", *cit.*, 2018, p. 979, 981.
[1308] DIAKOPOULOS, N., "Accountability in algorithmic decision making", *cit.*, 2016, p. 58.
[1309] Article 41 of the European Charter of Fundamental Rights.
[1310] BRKAN, M., "Do algorithms rule the world?...", *cit.*, 2017, pp. 23-24.
[1311] MALGIERI, G. & COMANDÉ, G., "Why a right to legibility of automated decision-making exists in the General Data Protection Regulation", *International Data Privacy Law*, vol. 7, No. 4, 2017b, p. 244.
[1312] *Idem*, pp. 260-261.
[1313] BINNS, R. *et al.*, "It's reducing a human being to a percentage; perceptions of justice in algorithmic decisions", paper presented at the *ACM Conference on Human Factors in Computing Systems (CHI'18)*, 2018, p. 4. Available on 10th May 2019 at: https://arxiv.org/

subject's case and 4) providing aggregate information regarding the results of individuals in the same demographic groups as the data subject.

While the suggested forms of explanation are not presented in a cumulative manner, materialising the analysed right by aggregating the four types of explanations described above would probably be the best way in which to make the rights to information, access and explanation effective. This will however not always be possible seeing as, in some cases, conflicts with trade secrets could arise. However, considering the fact that revealing the source code to data processing tools will not provide individuals with information that they can actually understand and therefore use in order to challenge profiles and decisions based on the processing of their personal data, focus should be on establishing a common ground in which individuals can obtain understandable explanations without harming any intellectual property rights.

WACHTER, MITTELSTADT and RUSSELL, offer an alternative possibility that relies on offering data subjects "counterfactual explanations" regarding the results of automated processing. These explanations would tell individuals what they should change in order to obtain a different result from the automated decision-making process. For example, if an individual did not get a loan she applied for, the explanation provided would perhaps indicate that her application was rejected because she did not have enough savings and establish the amount of savings that would have been necessary in order for her loan application to be accepted.[1314] They argue that this would offer individuals with a much clearer picture of the elements that really matter to them without entering conflicts with trade secrets or other intellectual property rights.[1315] Additionally, the implementation of this system would partly satisfy and be compatible with proposals regarding the possibility that individuals experiment with the models used in order to make automated decisions that concern them.[1316]

In order to solve the conflicts that arise regarding the rights to information, access and explanation of data subjects more specific regulatory instruments should be developed, ideally at the European level, although more likely at the member state level, assessing the degree of disclosure that should be provided in different scenarios. The proposals that are

[1314] WACHTER, S., MITTELSTADT, B. D. & RUSSELL, C., "Counterfactual explanations without opening the black box…", *cit.*, 2018, pp. 844-845.
[1315] *Idem,* p. 871.
[1316] WACHTER, S., MITTELSTADT, B. D. & RUSSELL, C., "Counterfactual explanations without opening the black box…", *cit.*, 2018, p. 880; CITRON, D. K. & PASQUALE, F., "The scored society…", *cit.*, 2014, p. 29.

indicated above should be considered and probably combined in order to offer appropriate explanations for each given scenario.

3.1.2. Information, access and explanation rights in Directive 2016/680 for data protection in law enforcement: the conflict with state and public security

A particularly complicated trade-off is the one that arises within law enforcement between public security and transparency. The degree to which individual fundamental rights can be restricted by law enforcement authorities and the criminal justice system is exactly what led to special safeguards that protect individuals, such as due process rights, being placed in criminal procedures. Amongst the rights that are recognised to individuals in most Western criminal law systems are information rights.[1317] Thus the recognition of access and information rights by Directive 2016/680 for data protection in law enforcement is consistent and reinforces the protection of individuals' fundamental rights through safeguards in procedural criminal law.

However, the threats to public security that can be caused in certain cases by the access to information of investigated or prosecuted individuals or groups can justify limiting transparency-related rights. Consequently, unlike the GDPR, Directive 2016/680 for data protection in law enforcement does not recognises a general principle of transparency for data processing that must rule over all personal data-related activities that fall within its scope of application.[1318] Consequently, and although Recital 38 does recognise a right of explanation as an expression of the right to human intervention in solely automated decision-making contained in article 11 of the Directive, said right to explanation, as well as the rights to information and access (articles 13 and 14 of the Directive) are much more limited than those contained in the GDPR.

Additionally, while articles 13, 14 and 15 of the GDPR indicate that "meaningful information about the logic involved" must be provided in solely automated decision-making, the Directive does not require said information to be provided to individuals, thereby limiting their rights to challenge decisions made by automated processes seeing as the justification obtained for decisions made will be very restricted. Moreover, it does not even require that

[1317] See, for example, Directive 2012/13/EU of the European Parliament and of the Council of 22 May 2012 on the right to information in criminal proceedings.
[1318] See articles 4.1(a) of Directive 2016/680 and 5.1(a) of the GDPR.

individuals be informed that solely automated decision-making, including profiling, is taking place.

Article 13.3 of the Directive establishes a set of cases in which the right to information can be limited and article 15 does the same with respect to the right of access. In both cases, authorities must ensure that "such a measure constitutes a necessary and proportionate measure in a democratic society with due regard for the fundamental rights and the legitimate interests of the natural person concerned". However, even if this means that restrictions to the rights to access and information can only be carried out when they are proportional and legitimate, all in all, the existence of a right to explanation in the Directive for data protection in law enforcement and the criminal justice system is rather doubtful.

The Directive is the clear materialisation of the conflicts between the rights to explanation, information and access and public interests. For example, when restrictions to the rights to information and access are carried out under the scope of article 15 of the Directive, said restrictive measures must always be taken in order to either to avoid the obstruction or prejudicing law enforcement activities or to protect public or national security or the rights and freedoms of others.

The grounds for exemption of the rights to information and access show the necessary limits to transparency that have to be set in some cases in order to safeguard other rights or interests. One of the reasons for the aforementioned restrictions is avoiding that individuals have the means to game the system by knowing in advance, for example, what elements draw red flags at airport security checks or when detecting fraud or tax evasion.[1319]

The tools offered to individuals through transparency-related rights in order to challenge decisions based solely or partly on automated processing are quite limited under the scope of Directive 2016/680 for data protection in law enforcement,[1320] which could lead to an excessive restraint to the due process rights of data subjects.

[1319] BAYAMLIOĞLU, E., "Transparency of automated decisions in the GDPR…", *cit.*, 2018, p. 18; ZARSKY, T., "Transparent predictions", *cit.*, 2013, pp. 1553-1554.
[1320] DIMITROVA, D. & DE HERT, P., "The right of access under the police directive: small steps forward" in MEDINA, M. *et al.*, (eds.) *6th Annual Privacy Forum 2018: Privacy Technologies and Policy*, Berlin, Springer, 2018, pp. 127.

3.1.3. A few final remarks with regard to the transparency principle and the rights that derive from it

The widespread concern with providing individuals with explanations on automated decision-making systems and profiling that they can understand and the demands for general algorithmic transparency have been met with criticisms by part of the scholarship. ZERILLI *et al.* claim that the expectations for algorithmic transparency result from the unrealistically high expectations created surrounding the transparency of human decision-making.[1321]

They argue that the explanations that are demanded that be provided with respect to automated decision-making are far more extensive than those required in human decision-making given that it is never possible to know exactly what the cognitive process behind a human-made decision is while this is exactly what we are asking of algorithms.[1322] Even if rational explanations for a decision are offered, it is always possible that there is an intrinsic discriminatory rationale behind human decision-making that will be less pervasive in algorithmic decision-making.[1323] Moreover, the explanation provided by humans may not even convey the real reasoning behind the decision. Therefore, they consider that, the explanations provided regarding automated systems should contain the reasoning behind the decision but not the architectural framework of the model employed in order to avoid clashes with intellectual property rights.[1324]

The convenience or lack thereof of including architectural explanations is however more related to systemic transparency and auditing and will therefore be covered in the following section. However, with respect to the issue at hand, that is, transparency-related rights, what ZERILLI *et al.* claim is actually in line with recent proposals put forward by the literature which try to provide a right to explanation without opening the "black box".[1325]

Being adequately informed about the logic behind algorithmic-generated decisions or profiles is necessary for the individual to justify any claim regarding the decision or profile.[1326] Consequently, since the right to explanation aims to provide individuals with enough tools to

[1321] ZERILLI, J. *et al.*, "Transparency in algorithmic and human decision-making…", *cit.*, 2018, p. 2.
[1322] *Idem*, p. 5.
[1323] *Idem*, p. 13.
[1324] *Idem*, p. 16.
[1325] See, for example, WACHTER, S., MITTELSTADT, B. D. & RUSSELL, C., "Counterfactual explanations without opening the black box…", *cit.*, 2018, pp. 841-887.
[1326] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1588.

defend themselves with regard to decisions that affect them, providing the reasoning behind the decision made is enough to contest it. This becomes especially relevant in administrative decision-making where justification of most administrative actions is made compulsory by the relevant laws in order to prove the fairness, legitimacy and validity of the decision and offer individuals tools to challenge it.

Transparency can be constructed, both as a governance instrument and as tool for the protection of individual rights. In this section it has been analysed from the perspective of the protection of individual rights. The rights derived from the general transparency principle aim towards the protection of privacy and thus shape a framework of protection against discrimination from the perspective of anti-classification, especially, in the cases of the GDPR and Directive 2016/680 for data protection in law enforcement, when rights derived from the transparency principle are applicable with regard to the special categories of data recognised under article 9 of the GDPR and 10 of the Directive.

The reason why the debate on the right to an explanation and the other transparency-related rights contained in the GDPR has gained so much traction, especially when the Regulation was first passed in 2016, is that these rights have been seen by part of the scholarship as a possible solution to the problems of opacity generated by machine learning systems. As indicated in the previous paragraph, if properly articulated, these rights could help to protect not only privacy-related rights but also other rights that can be affected by automated processing, such as the right to be heard or the rights to equality and non-discrimination. However, the significant limitations to transparency-related rights that result from the way in which they are structured in the GDPR render void the possibility of using these as mechanisms to achieve algorithmic transparency, at least for the time being.

3.2. THE RIGHT TO BE HEARD AND CONTEST DECISIONS: THE RIGHT TO AN EFFECTIVE REMEDY

In addition to transparency related rights, technological due process proposals also highlight the importance that individuals have the possibility of being heard when their data is being processed as well as being able to challenge decisions that affect them and which are reached through algorithmic processing of their data.[1327] The right to be heard as it has been presented in the literature includes the possibility of correcting collected or processed data that is not

[1327] *Idem*, p. 1555.

accurate.[1328] Additionally, rights that provide individuals with the possibility of challenging decisions before an impartial third party are also necessary in order to guarantee full due process rights.[1329]

CITRON also indicates that, in order to reduce "automation bias", individuals with the power to hear requests and complaints regarding any stage of algorithmic decision-making should be trained to be aware of the biases and mistakes that can result from automated data processing and should also provide justification for following the decision made by the automated system.[1330] She also argues for the intervention of expert testimony in hearings in order to assess any possible flaws of the algorithm.[1331]

The rights to be heard and contest semi-automated or automated decisions are an expression of the fundamental right to an effective remedy, recognised in article 47 of the Charter of Fundamental Rights of the EU and article 13 of the European Convention on Human Rights. In the context of public administrations, they are also an expression of the right to good administration contained in article 41 of the Charter of Fundamental Rights of the EU.

3.2.1. The right to be heard and contest decisions in the GDPR

3.2.1.1. Data subjects' due process rights in art. 22

Article 22 of the GDPR establishes a minimum set of rights that must be guaranteed to data subjects when the exceptions to the general prohibition of being subjected to decisions based solely on automated processing, including profiling, apply. These provisions have been interpreted to be "due process" rights for individuals being subjected to algorithmic decision-making.[1332] These due process rights include the data subject's right to obtain human intervention, to express his or her point of view and to challenge the decision as well as the right to explanation that was already analysed in the previous section. The rights that will be analysed in the following pages represent the minimum safeguard threshold and are therefore meant to be specified and further developed by member state legislation.

[1328] CRAWFORD, K. & SCHULTZ, J., "Big data and due process…", *cit.*, 2014, p. 127.
[1329] *Idem*, pp. 127-128.
[1330] CITRON, D. K, "Technological due process", *cit.*, 2008, p. 1306.
[1331] *Ibidem*.
[1332] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1592-1593.

i) *The right to obtain human intervention*

The degree of human intervention that would be considered acceptable to satisfy this right is unclear.[1333] The A29WP Guidelines, although quite vague on this issue, do indicate that "any review must be carried out by someone who has the appropriate authority and capability to change the decision"[1334] and it is expected that member states will develop more specific regulations regarding this issue.

As it was already discussed when addressing the scope of article 22, it is not clear that human involvement does in fact provide safeguards to prevent discrimination due to "automation bias", [1335] and prejudices held by humans towards traditionally oppressed groups. Additionally, if not properly regulated this possibility could lead to the person who is involved in the automated decision-making process to serving as a scapegoat on whom to blame discriminatory, erroneous or unfair outcomes that are actually the result of the whole system of algorithm creation and deployment.[1336]

These possible downsides to the right to obtain human intervention should however not be misunderstood as a defence of withdrawing all possibility of requesting human intervention. If correctly applied, this right ensures that some supervision will take place over the algorithms making decisions or drawing up profiles on the individual. Member state regulations should therefore consider the issues mentioned when setting up regulations concerning human involvement with regard to article 22 ensuring that the right to request and obtain human intervention is sufficiently protected and that the appropriate measures to guarantee its effectiveness are set.

ii) *The right to express his or her point of view*

The data subject's right to express his or her point of view has been analysed in close relation with the rest of the minimal safeguards established by article 22.3 of the GDPR as well as with the rights to information, access and explanation, which are regulated by articles 13 to 15 of the GDPR and which are also set up as minimal safeguards, equivalent to the ones contained in article 22.3 by Recital 71. In this sense, only if the data subject has access to the

[1333] *Idem*, p. 1594.
[1334] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", *cit.*, 2018, p. 27.
[1335] CITRON, D. K, "Technological due process", *cit.*, 2008, pp. 1271-1272.
[1336] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1594.

information used by the system and logic involved in the decision, can she express her opinion on the justification provided by the system and know what kind of information or statement to provide in order to defend her position.

Additionally, this right is very closely related to the right recognised in article 16 of the GDPR, which states the following:

> "The data subject shall have the right to obtain from the controller without undue delay the rectification of inaccurate personal data concerning him or her. Taking into account the purposes of the processing, the data subject shall have the right to have incomplete personal data completed, including by means of providing a supplementary statement".

This is precisely one of the ways in which the data subject's right to express her opinion could be made effective. The combination of these two rights is essential in order to prevent algorithmic-based discrimination from occurring given the fact that, as it was stated in Part I, the vast amounts of data that machine learning algorithms can sometimes analyse leads to missing information or errors in the data, which are nevertheless tolerated given the fact that the computational capacity of machine learning algorithms enable them to reach generally accurate results even when some errors are contained in the dataset.

Both safeguards analysed result especially useful with regard to how data protection regulations have been insofar applied in certain member states. For example, in Spain, data protection regulations have generally favoured the publicity of negative financial information, such as debtor registries, while implementing constraints with regard to access to positive financial information, for which the data subject's authorisation is generally needed. Thus financial data profiles are generally structured as negative profiles rather than including the whole picture regarding individuals' financial situation.[1337] The rights to express the data subject's point of view and to rectification should provide data subjects with a way in which to force controllers and processors to complete their profiles and increase their accuracy.

[1337] MAS BADÍA, M. D., "Los ficheros de solvencia patrimonial en la proyectada nueva Ley Orgánica de Protección de Datos de carácter personal. ¿Un avance o una oportunidad perdida?", *Actualidad Civil*, No. 11, 2017, p. 93; PALMA ORTIGOSA, A., "Decisiones automatizadas en el RGPD. El uso de algoritmos en el contexto de la protección de datos", *Revista General de Derecho Administrativo*, No. 50, 2019, p. 33.

iii) *The right to challenge the decision*

The right to challenge the decision is the possibility the individual has to contest the decision before the controller and processor.

While the right to challenge the decision is included as a necessary safeguard for those cases in which the exceptions for the general prohibitions of article 22.1 and 22.4 apply, it is important to highlight that articles 77 to 79 of the GDPR also include the possibility of lodging complaints with the competent supervisory authority, of challenging the authority's decision before a court and, in general, of obtaining judicial remedies against controllers or processors. The main difference between said rights and the one recognised in article 22.3 is the fact that the latter is meant to provide the data subject with an additional (previous) step to presenting claims before a supervisory authority or a court.

Once again, the importance of having information available on the reasons why the decision was made is vital in order for the data subject to challenge it.[1338]

3.2.1.2. Individual rights to be heard and challenge decisions recognised outside of article 22 of the GDPR

The rights recognised under or with regard to the situations covered by article 22 of the GDPR have a limited scope of applicability. Even though human intervention must be relevant in order for the data processing and/or profiling to not fall under the scope of article 22, there will be many cases in which such reinforced protection does not apply. Moreover, even if the scope of applicability of article 22 is considered to be extensive enough, analysing the other rights recognised to all data subjects irrespective of the type of processing their data is subject to, is vital in order to illustrate the full scope of protections from the perspective of individual rights offered by the GDPR.

i) *The rights to data portability, rectification, erasure and restriction of processing*

The rights to rectification and erasure are expressions of the informational self-determination perspective in the sense that they provide individuals with the possibility of protecting

[1338] TODOLÍ SIGNES, A., "La gobernanza colectiva de la protección de datos en las relaciones laborales: big data, creación de perfiles, decisiones empresariales automatizadas y los derechos colectivos", *Revista de Derecho Social*, 2018, No. 84, p. 80.

themselves against inaccuracies or incomplete datasets. More importantly, by exercising the right to erasure, individuals can, in certain cases, request the elimination of pieces of their personal data that have been made available to the public.

The right to rectification contained in article 16 of the GDPR, which was already addressed, offers the data subject the opportunity of correcting inaccurate information and completing information regarding both input and output data.[1339]

The right to erasure (right to be forgotten) contemplated in article 17 provides individuals with the possibility of eliminating input or output[1340] personal data when it no longer serves the purposes it was collected or processed for; when the data subject withdraws consent and there is no other legal ground for processing; when the data have been unlawfully processed; when erasure is compulsory with regard to Union or member state law and when personal data have been collected in relation to the offer made to a minor under 16 of information society services.

This right has been widely discussed due to the series of CJEU Judgments regarding its application by Google in search results associated to individuals' names.[1341] While there are many elements that can be discussed, what becomes particularly relevant is the fact that the CJEU recognises ample powers to search engine operators when making this right effective. Search engine operators are responsible for carrying out the balancing test of the different interests at hand, namely the data protection rights of applicants and the general right to information.[1342]

Personal data erasure can also be requested when exercising the right to object recognised in article 21 of the GDPR, which will be analysed later on. In these cases, the controller has to eliminate individuals' data, as long as they request said erasure, when they object to the processing, including profiling, necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or necessary for

[1339] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", *cit.*, 2018, p. 18.
[1340] *Ibidem.*
[1341] CJEU Judgments 13th May 2014, C- 131/12 Google Spain SL and Google Inc. v. Agencia Española de Protección de Datos (AEPD) and Mario Costeja González; 24th September 2019, C-136/17, GC and Others v. Commission nationale de l'informatique et des libertés (CNIL) and C-507/17, Google LLC, successor in law to Google Inc. v. Commission nationale de l'informatique et des libertés (CNIL).
[1342] BOIX PALOP, A., "El equilibro entre los derechos del artículo 18 de la Constitución, el «derecho al olvido» y las libertades informativas tras la Sentencia Google", *Revista General de Derecho Administrativo*, No. 38, 2015.

the legitimate interests pursued by the controller or by a third party. This obligation could however be overridden if legitimate grounds for the processing take place. Nevertheless, if the data subject exercises the right to object to the processing of their personal data, including profiling, for direct market purposes, the controller will, with no exceptions, be forced to comply with the erasure of the individual's personal data if she requests it.

When personal data has been shared, article 17 of the GDPR states that "the controller, taking account of available technology and the cost of implementation, shall take reasonable steps, including technical measures, to inform controllers which are processing the personal data that the data subject has requested the erasure by such controllers of any links to, or copy or replication of, those personal data".

A concern pointed out with regard to these rights is the fact that individuals may use them to their advantage in order to game the system, eliminating information that is actually accurate.[1343] Another related concern that will be further discussed later on is the fact that these rights will probably be mainly exercised by individuals with more resources, thereby introducing systemic biases in automated decision-making structures and further enhance inequalities.[1344]

Article 18 of the GDPR, recognises the data subject's right to obtain from the controller restriction of processing at any stage of data processing, including profiling in certain cases[1345] and is mainly designed as an alternative protection mechanism to the right to erasure, an additional safeguard that data subjects can claim in the interim between the time in which the rights to rectification and to object are exercised and their effects take place.

[1343] BAMBAUER, J. & ZARSKY, T., "The algorithm game", *Notre Dame Law Review*, vol. 94, No. 1, 2018, p. 35.

[1344] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1581: "A system of individual rights can conflict with system-wide accuracy and system-wide concerns about bias. Individuals could use correction and erasure rights to intentionally game a decision-making algorithm, by changing or eliminating negative information about themselves. Even just opting out of a system without actively introducing inaccuracies can affect system bias. For example, if only well-educated, socioeconomically elite individuals opt out, then the contours of a machine learning system would come to reflect those who are less empowered and remain within it. Allowing individuals to alter or erase their information changes the data set, which changes the algorithm going forward. There are thus real tensions between attending to individual dignitary or autonomy concerns on the one hand, and addressing systemic bias and discrimination on the other."

[1345] The right to restriction of processing shall apply during the time which the controller is verifying that data is accurate when said accuracy has been contested by the data subject; when the processing is unlawful but the data subject chooses to exercise this right instead of the right to erasure; when the controller no longer needs the personal data for the purposes of the processing, but they are required by the data subject for the establishment, exercise or defence of legal claims; and when the data subject has exercised the right to object during the time in which the existence of legitimate grounds for processing is being verified in the case that they could apply (article 18 of the GDPR).

The rights to rectification, erasure and restriction of processing are further enhanced and protected by the right of notification recognised in article 19 of the GDPR which forces controllers to communicate any rectification or erasure of personal data or restriction of processing to each recipient[1346] to whom the personal data have been disclosed, unless this proves impossible or involves disproportionate effort.

Finally, the right to data portability is included in article 20 of the GDPR and it constitutes one of the materialisations of the individual informational self-determination approach that some articles in the Regulation follow. Providing individuals with the possibility of taking their personal data from one controller to another helps to level the playing field between data subject and controller, as she can choose to remove her data from the controller and take it elsewhere.[1347]

ii) *The right to object*

Article 21 of the GDPR provides individuals the right to object to the processing of their information, including profiling, "on grounds relating to his or her particular situation" in certain cases[1348] which, if exercised, leads to the controller's obligation to stop said processing unless compelling legitimate grounds for the processing which override their data subject's rights are demonstrated. Legitimate grounds for processing do however not serve as a valid reason for overriding an individual's right to object if said right is exercised to object to the processing of their personal data, including profiling, for direct market purposes, which means that the controller will, with no exceptions, be forced to stop processing her data.

iii)*The rights to lodge complaints before supervisory authorities and to judicial remedies*

Articles 77 to 79 of the GDPR recognise individuals' rights to "lodge a complaint with a supervisory authority [...] if the data subject considers that the processing of personal data relating to him or her infringes" the GDPR (article 77); "to an effective judicial remedy against a legally binding decision of a supervisory authority concerning them" (article 78);

[1346] Article 4.9 of the GDPR: "A natural or legal person, public authority, agency or another body, to which the personal data are disclosed, whether a third party or not. However, public authorities which may receive personal data in the framework of a particular inquiry in accordance with Union or member state law shall not be regarded as recipients; the processing of those data by those public authorities shall be in compliance with the applicable data protection rules according to the purposes of the processing".
[1347] MAZUR, J., "Automated decision-making and the precautionary principle in EU law", *cit.*, p. 9.
[1348] When processing, including profiling, is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or necessary for the legitimate interests pursued by the controller or by a third party.

and to "an effective judicial remedy where he or she considers that his or her rights under [the GDPR] have been infringed as a result of the processing of his or her personal data in non-compliance" the Regulation (article 79). In addition, article 80 establishes the right that data subjects have to confer their representation in the previously stated cases to non-profit organisations which are active in the "protection of data subjects' rights and freedoms with regard to the protection of their personal data". Article 82 recognises data subject's right to compensation.

3.2.2. The rights to be heard and challenge decisions in Directive 680/2016

As it happens with the rest of the Directive, the active rights recognised to individuals mainly follow the same structure as the GDPR. For example, article 11 of the Directive recognises the right of individuals to obtain human intervention when they are subjected to decision-making based solely on automated processing, including profiling and Recital 38 of the Directive indicates that the rights to express the data subject's point of view and to challenge decisions are understood to be comprised under said right to obtain human intervention.

Additionally, article 16 recognises the right to rectification, erasure and restriction of processing. The controller shall restrict processing instead of erasing the data are when "the accuracy of the personal data is contested by the data subject and their accuracy or inaccuracy cannot be ascertained or the personal data must be maintained for the purposes of evidence".

The exercise of these rights also generates transparency-related obligations for controllers who must inform individuals when they refuse to correct or erase information or restrict processing of data. However, these obligations can be exempted in order to avoid the obstruction or prejudicing law enforcement activities or to protect public or national security or the rights and freedoms of others.

Under the Directive data subjects do not have a similar right to the one contained in article 21 of the GDPR which allows them to object to the processing of their data. Finally, in the same way as the GDPR does in articles 77 to 79, the Directive recognises in articles 52 to 54 the rights to lodge a complaint before the supervisory authority, to an effective judicial remedy against a supervisory authority, a controller or a processor.

3.2.3. Positive and negative aspects of the due process rights system contained in the GDPR

The individual rights here analysed offer a series of solutions towards the general concerns that were pointed out in the previous chapter with regard to the need to the expansion of algorithmic decision-making. Generally unfair and, more specifically, discriminatory outcomes, can be prevented and dealt with through active rights that allow data subjects to claim that the data used is not accurate or that the decision made is based on prohibited grounds. In addition, these rights also address dignitary and autonomy related concerns by offering data subjects the possibility of adopting an active role in the processing of their personal data, thereby acknowledging their humanity and agency. Finally, individual transparency rights is essential in order for data subject's to acquire the necessary knowledge to challenge automated decisions. It is also worth noting that while the catalogue of rights available is more expansive when solely automated processing takes place, there are also many other rights that can be exercised by data subjects in order to intervene and prevent wrongful, unfair or discriminatory data processing and decision-making, when these actions do not fall within the scope of article 22 of the GDPR.

The GDPR is mostly built around the idea of informational self-determination, which provides individuals with the possibility of actively deciding how and if their data is processed. The informational self-determination approach could be expanded even further than the catalogue of rights available to data subjects in the GDPR. However, considering the risks generated by machine learning algorithms, opt-in systems would probably be better in order to ensure that individuals made more conscious choices of the information that they share.

In addition, while much of the scholarship has focused on the scope of the "right to an explanation", the other rights contained in the GDPR, such as the right to erasure or restriction of processing provide tools for individuals to exercise in cases in which obtaining an explanation and challenging the decision is simply not enough. For example, when a piece of wrong information regarding an individual appears in search engine results, said individual

will generally not be concerned about how the decision to post said information was reached but will simply want for it to disappear.[1349]

4. REGULATORY MECHANISMS FOR SYSTEM TRANSPARENCY AND ACCOUNTABILITY THROUGH DATA PROTECTION

4.1. REGULATORY FRAMEWORKS

In order to set up a series of regulations and/or norms that render algorithms accountable, it is necessary to reflect on the different possible regulatory frameworks that could be implemented for the private use of automated decision-making. The following pages briefly explain how self-regulation, co-regulation and state intervention mechanisms could operate in order to set up a framework to ensure algorithmic fairness and accountability.

4.1.1. Self-regulation

Self-regulation can take place internally within a firm or within a certain sector of activity. These initiatives, when adopted by particular firms, are generally framed within their corporate social responsibility strategy and are aimed towards improving their reputation.[1350] This means that there generally has to be an incentive, such as pressure from customers,[1351] for organisations to self-regulate and adopt measures, such as teaching ethics to programmers, developing diversity hiring strategies or introducing review mechanisms to ensure their automated systems do not generate discrimination.

For example, some companies, such as Facebook, have announced the creation and implementation of software tools for detecting biases in the machine learning systems used by the company.[1352] Additionally, many of the big tech giants have also developed general principles[1353] and responsible AI practices[1354] and report on their workforce's diversity.[1355]

[1349] EDWARDS, L. & VEALE, M., "Slave to the algorithm? …", *cit.*, 2017, p. 41.
[1350] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, p. 41.
[1351] *Ibidem.*
[1352] KLONICK, K., "The Facebook oversight board: creating an independent institution to adjudicate online free expression", *Yale Law Journal*, vol. 129, No. 8, 2020, pp. 2418-2499; TEICH, P., "Artificial intelligence can reinforce bias, cloud giants announce tools for AI fairness", *Forbes*, 24th September 2018. Available on 19th September 2019: https://www.forbes.com/
[1353] GOOGLE AI, "Artificial intelligence at Google: Our principles", 2019. Available on 19th September 2019 at: https://ai.google/principles/
[1354] GOOGLE AI, "Artificial intelligence at Google: Responsible AI practices", 2019. Available on 19th September 2019 at: https://ai.google/principles/

However, the companies that address these issues are those that are more visible as a consequence of directly interacting with consumers. Hence, it is less likely that firms which provide services to other businesses have incentives to place restrictions on their activities in order to ensure that the algorithms they develop and sell are designed in a way that adequately protects the right to equality and non-discrimination.[1356]

Additionally, the current state of technological development prevents consumers from fully comprehending or even being aware of the implications that automated systems have, meaning that the likelihood of demands for ethical, non-discriminatory algorithms is severely reduced.[1357]

Sector self-regulation mechanisms can be more effective in governing algorithmic discrimination seeing as the standards set can reach every firm in the sector and not only those firms that are more visible and that therefore work harder to protect their reputation. Amongst the different mechanisms that can be adopted by economic sectors to self-regulate are codes of conduct, technical standards, organisational standards and certification mechanisms which can sometimes include quality seals.[1358] In order to set those mechanisms or review the actions taken by industry members, review or dispute resolution boards and ethics committees can also be set up.[1359]

While sector or industry self-regulation is an important step with regard to raising awareness of the risks that data processing can generate and implementing mechanisms to prevent and deal with said risks,[1360] relying on this form of governance framework alone is insufficient, particularly when the objective is to establish regulations that protect against the risks that algorithms pose on fundamental rights. Self-regulation unavoidably favours sector interests over concerns regarding the protection of equality, non-discrimination and other human rights, which means that, regardless of the standards, codes of conduct and other self-

---

[1355] GOOGLE, "Google diversity annual report 2018", 2018, p. 18.; MCINTYRE, L., "Diversity and inclusion update…", *cit.*, 2018; WILLIAMS, M., "Facebook 2018 diversity report…", *cit.*, 2018; APPLE, "Inclusion and diversity", *cit.*, 2017.
[1356] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, p. 41.
[1357] *Idem*, p. 40.
[1358] O'NEIL RISK CONSULTING AND ALGORITHMIC AUDITING (ORCAA), "Services", 2019. Available on 19th September 2019 at: http://www.oneilrisk.com
[1359] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, p. 41.
[1360] CATH, C. *et al.*, "Artificial intelligence and the 'good society'…", *cit.*, 2018, p. 513.

regulatory mechanisms that might be developed by certain sectors, public institutions develop a more comprehensive and direct system to tackle issues of algorithmic discrimination.[1361]

4.1.2. Co-regulation (or regulated self-regulation)

Another form of regulation that involves industry members is what can be labelled as co-regulation[1362] or regulated self-regulation.[1363] Co-regulation entails the cooperation between public bodies and private firms in setting up and enforcing regulatory instruments and mechanisms.[1364] There are different ways in which co-regulatory frameworks can be developed. In some cases, public bodies set the objectives that industries must comply with and the latter are free to implement and enforce regulatory instruments as they see fit[1365] while, in other cases, rules and standards are developed by industry members while public bodies offer tools and mechanisms for supervision and enforcement.[1366]

Some of the mechanisms used in self-regulation and co-regulation overlap but differ in the fact that when they are used within a self-regulatory governance framework there is no intervention by public bodies. For instance, there can be a purely private certification scheme that is not mandated or recommended by any regulatory instruments but that private firms choose to subject their algorithms to in order to provide them with a greater degree of legitimacy.

It is finally necessary to point out that co-regulatory regimes can work alongside command-and-control instruments that are generally framed within state intervention regulatory regimes.[1367] The implementation of more interventionist measures is necessary in order to guarantee the enforcement of softer law mechanisms such as codes of conduct and avoid the risk of regulatory capture that frameworks of regulatory collaboration between the public and private sector can produce.[1368]

[1361] *Ibidem*.
[1362] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1599-1607; KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, pp. 43-45.
[1363] DARNACULLETA I GARDELLA, M., *Autorregulación y Derecho Público: La Autorregulación Regulada*, Madrid, Marcial Pons, 2005.
[1364] HIRSCH, D. D., "The law and policy of online privacy: Regulation, self-regulation, or co-regulation", Seattle University Law Review, vol. 34, No. 2, 2011, p. 441.
[1365] MARSDEN, C. T., *Internet Co-regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace*, Cambridge, Cambridge University Press, 2011, p. 54.
[1366] HIRSCH, D. D., "The law and policy of online privacy…", *cit.*, 2011, p. 442.
[1367] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1564.
[1368] *Idem*, p. 1566.

4.1.3. Regulation (state intervention)

The final form of governance framework is regulation or state intervention. There are different types of mechanisms that fall under this framework. These mechanisms range from the more direct and restrictive prohibition measures to lighter forms of regulation that simply nudge firms that develop algorithms and tech experts so that they introduce the necessary mechanisms to ensure that the risks caused by algorithms to individual's fundamental rights are minimised.[1369]

Some of the main arguments that are wielded against state intervention is the risk that imposing very restrictive regulations on the use of algorithms might lead to important setbacks in the development of technology. If the costs associated to the development of algorithms become very high due to the requirements set by regulatory instruments, the incentives to invest in said technology could disappear, investment could be reduced and the social and economic benefits that could be yielded from the development of automated systems mey eventually be lost.[1370]

However, state intervention is necessary in many instances in order to ensure that innovation is directed in a certain way.[1371] Without a regulatory system that promotes and, in many cases, forces firms to embed fundamental right protection as a characteristic of automated systems, it is very unlikely that algorithm developers will choose to find innovative ways in which to introduce said requirements.

4.2. THE GDPR AS A SYSTEM OF GOVERNANCE

As a significant part of the scholarship[1372] has accurately pointed out, the GDPR does not only build a system of algorithmic regulation through individual rights and very specific prohibitions and mandates but also contains a series of provisions through which algorithms can be generally controlled. These provisions are what KAMINSKI labels the "collaborative governance regime"[1373] in the GDPR. The aforementioned label responds to the fact that the

[1369] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, pp. 48-49.
[1370] YEUNG, K. & LODGE, M., "Algorithmic regulation…", *cit.*, 2019, p. 7.
[1371] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, p. 46.
[1372] See, for example, EDWARDS, L. & VEALE, M., "Slave to the algorithm?…", *cit.*, 2017, pp. 18-84; KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1529-1616.
[1373] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1563.

tools set up for these purposes in the GDPR are mainly set up from the perspective of co-regulatory governance.

However, as it was indicated in the previous section, self-regulation and co-regulation mechanisms can overlap, hence the lines between self-regulation and co-regulation do become quite blurry in some of the GDPR's provisions. A clear example of this is the fact that, as a consequence of the way in which certification mechanisms are regulated in the GDPR, part of the literature has identified them as self-regulation mechanisms[1374] while others analyse them within the framework of co-regulation.[1375] In addition, the GDPR's "collaborative governance regime"[1376] also contains some elements that approximate it more to a hard law regime as they impose direct obligations on data processors and, especially, controllers.[1377]

The mechanisms here analysed are those which the GDPR includes in order to ensure that "data protection by design and by default"[1378] is incorporated to automated or semi-automated processing systems. These tools, amongst which codes of conduct and certification mechanisms are included, are not only analysed from the perspective of the GDPR and data protection, but are also addressed in a more general manner as elements that can provide system transparency and accountability.

4.3. SYSTEM TRANSPARENCY AND ACCOUNTABILITY

Much of the scholarship concerned with providing oversight tools for automated or semi-automated systems has approached system transparency and accountability as two elements that are dependent on one another.[1379] In fact, since the point of departure when analysing system transparency and accountability is the idea that the former will provide the latter, much of the debate has been focused exclusively on opening "the black box".[1380] Revealing the logic behind a decision in particular cases and the way in which an automated system generally works is important to a certain extent in order to examine the system and control it

[1374] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, p. 43.
[1375] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1599-1600.
[1376] *Idem*, p. 1563.
[1377] *Idem*, p. 1597.
[1378] See article 25 of the GDPR and 20 of Directive 2016/680.
[1379] KROLL, J. *et al.*, "Accountable algorithms", *cit.*, 2017, pp. 657-658; EDWARDS, L. & VEALE, M., "Slave to the algorithm?…", *cit.*, 2017, p. 41.
[1380] PASQUALE, F., *The Black Box Society…*, *cit.*, 2015.

but, even when necessary for accountability, transparency is only the first stepping stone in ensuring algorithms can undergo proper oversight and control.[1381]

While the "inherent goodness" of transparency can be argued,[1382] it is probably best to focus on the possibility of rendering such systems accountable rather than exclusively focusing on making systems transparent and hope that said transparency indirectly results in algorithmic accountability. Moreover, focusing on transparency as the only means through which to achieve algorithmic accountability can lead to ignoring other possible algorithmic oversight and control solutions. This is especially problematic when system transparency cannot be achieved or is not useful either due to the need to keep certain elements opaque to avoid gaming and the revelation of proprietary information or because the type of algorithm used cannot be held accountable even when its parameters and logic are disclosed.[1383]

It is useful to draw a comparison between aspirations and actual effectiveness of algorithmic transparency and the transparency mandate that democratic governments and public administrations are subjected to. Legislative instruments directed towards providing better and more transparent public powers have generally fallen short in achieving said objectives not only due to the capacity that public institutions have in controlling when and how information is disclosed but also, and largely, due to the massive amounts of data released, which are sometimes incomprehensible and overwhelm those citizens and organisations that may want to act as watchdogs.[1384] In order to have a full picture of the way in which an automated system works, the information provided should comprehend the instructions that the algorithm follows, the way in which it was trained, the data used to train it and the data it is fed when making decisions. This vast amount of information, which in many instances will also be difficult to understand and will be constantly updated, will likely hinder the possibility of holding algorithms accountable through transparency.

As the following chapter will convey, the need to focus on making systems accountable does not mean that full transparency should not be required for it is a necessary prerequisite to ensure accountability. Mechanisms to provide system transparency are necessary, particularly when algorithms are being used by the public sector, however, a wide array of tools that focus equally on transparency and accountability and, in some occasions, more on the latter

[1381] EDWARDS, L. & VEALE, M., "Slave to the algorithm?…", *cit.*, 2017, p. 41.
[1382] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, pp. 1018-1119.
[1383] KROLL, J. *et al.*, "Accountable algorithms", *cit.*, 2017, pp. 658-659.
[1384] EDWARDS, L. & VEALE, M., "Slave to the algorithm?…", *cit.*, 2017, p. 41.

than on the former, must also be developed in order to ensure better oversight and control of algorithmic systems.

Each of the tools analysed in the following pages must be understood as part of a comprehensive co-regulatory system of algorithmic accountability. Each mechanism serves different purposes and will therefore operate at different stages in the process of governing algorithms, whether it is when setting rules that algorithms must comply with (safe harbour agreements, codes of conduct and standards), when controlling that algorithms comply with said rules (impact assessments, auditing and certification) or aimed towards enforcing mandatory provisions (committees, authorities, boards and sanctions).[1385]

4.4. REGULATORY TOOLS FOR SYSTEM TRANSPARENCY AND ACCOUNTABILITY

4.4.1. Rule-setting mechanisms

4.4.1.1. Safe harbour and privacy shields

Before addressing the instruments contained in the GDPR it is relevant to briefly refer to the failed experience of establishing a co-regulatory framework for privacy protection between the EU and the US.

The safe harbour data protection agreement between the US and EU of 2000 established a series of principles that US firms had to comply with in order for transatlantic data transfers to take place. Private firms controlled compliance with the agreement's data privacy principles.[1386] However, as it was later found out, said controls were largely ineffective and the principles set in the agreement were not being respected. In 2015, the CJEU determined that the agreement was invalid as it did not establish enough safeguards to ensure the protection of EU citizens' personal data.[1387] Following said judgment, the EU Commission adopted the EU-US Privacy Shield Decision,[1388] which was recently invalidated by the

[1385] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1568-1569.
[1386] WEISS, M. A., & ARCHICK, K., "US-EU data privacy: from safe harbor to privacy shield", *Congressional Research Service*, 19th May 2016, pp. 5-6.
[1387] CJEU Judgment 6th October 2015, C-362/14, Maximillian Schrems v. Data Protection Commissioner and Digital Rights Ireland Ltd.
[1388] Decision (EU) 2016/1250 of 12 July 2016 pursuant to Directive 95/46/EC of the European Parliament and of the Council on the adequacy of the protection provided by the EU-U.S. Privacy Shield (notified under document C(2016) 4176).

CJEU[1389] as the Court determined that the Privacy Shield did not offer EU citizens sufficient protection, for US public authorities have large powers in accessing personal data that are in the possession of any company based in the US.

4.4.1.2. Codes of conduct

The development of algorithms is not framed within a regulated profession such as architecture or law,[1390] which means that codes of conduct are not generally enforceable. Moreover, the software systems that may create risks of discriminating and reinforcing traditional structures of oppression can be developed by individuals that do not necessarily fit into traditional definitions of certain professions for they have not had specific or regulated training.[1391] While corporations can choose to adopt codes of conduct set by some external organisation that is representative of the sector's interests, the degree to which said codes are actually applied may vary widely between firms.[1392] Additionally, these instruments generally contain very general principles that cannot be easily translated into exercisable mandates and that can be interpreted by algorithm developers in many different ways.[1393]

A self-regulatory governance framework might actually achieve a certain degree of effectiveness if software industry organisations such as the IEEE[1394] (Institute of Electrical and Electronics Engineers) set up enforcement boards that ensure compliance on the part of members. However, the effectiveness of said organisations will largely depend on the number of firms that decide to become members. While the use of this particular mechanism alone might not be sufficient to deal with the risks that the use of algorithms generates, including discrimination, it can be useful to set up the blueprint and inform technological firms as well as consumers of what general good practices should be carried out and what principles should be respected in the development and deployment of algorithms.

The GDPR includes several provisions which establish the possibility and need to encourage the development of codes of conduct. While the use of this compliance mechanism is referred to in different parts of the Regulation, it is article 40 that specifically regulates the use of

[1389] CJEU Judgment 16th July 2020, C-311/18, Data Protection Commissioner v. Facebook Ireland Ltd. and Maximillian Schrems.
[1390] BODDINGTON, P., *Towards a Code of Ethics for Artificial Intelligence*, Oxford, Springer, 2017, p. 59.
[1391] *Ibidem.*
[1392] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, p. 42.
[1393] *Ibidem.*
[1394] IEEE. Available on 2nd October 2019 at: https://www.ieee.org

codes of conduct by controllers in order to ensure that their activities are carried out according to the provisions of the GDPR. As some of the other governance mechanisms that are introduced by the GDPR, controllers and processors do not have an obligation to adopt codes of conduct but rather the GDPR indicates that "member states, the supervisory authorities, the Board and the Commission shall encourage the drawing up of said codes of conduct."

However, the cited provision does establish some general recommendations regarding the elements that should be considered by codes of conduct. Article 40.2 establishes a list of elements[1395] regarding which "associations and other bodies representing categories of controllers or processors may prepare codes of conduct, or amend or extend such codes." Those elements can therefore be interpreted as the recommended content for codes of conduct.

Additionally, there are also other elements introduced in article 40 of the GDPR that aim to guarantee a minimum degree of effectiveness and enforceability of these codes. Before a code of conduct is adopted, public supervisory authorities must examine it and indicate whether it complies with the GDPR (article 40.5). Codes of conduct are also to be published and, when a "code relates to processing activities in several member states", the European Data Protection Board must also review its content and provide its opinion on whether the code complies with the provisions of the GDPR before it is approved and published by the competent national Supervisory Authority (article 40.6 and 7).

All of the above mentioned elements are clearly indicative of how the codes of conduct, as regulated by the GDPR are more an instrument for co-regulation than self-regulation. Moreover, the way in which article 40. 1 delegates on member states (as well as other institutions and bodies) to encourage the drawing up and adoption of codes of conduct clearly shows how the GDPR is, in fact, structured more as a Directive than as a Regulation, thereby

[1395] Article 40.2 GDPR: "(a) fair and transparent processing; (b) the legitimate interests pursued by controllers in specific contexts; (c) the collection of personal data; (d) the pseudonymisation of personal data; (e) the information provided to the public and to data subjects; (f) the exercise of the rights of data subjects; (g) the information provided to, and the protection of, children, and the manner in which the consent of the holders of parental responsibility over children is to be obtained; (h) the measures and procedures referred to in Articles 24 and 25 and the measures to ensure security of processing referred to in Article 32; (i) the notification of personal data breaches to supervisory authorities and the communication of such personal data breaches to data subjects; (j) the transfer of personal data to third countries or international organisations; or (k) out-of-court proceedings and other dispute resolution procedures for resolving disputes between controllers and data subjects with regard to processing, without prejudice to the rights of data subjects pursuant to Articles 77 and 79."

enhancing the interpretation of the Data Protection Regulation as a collaborative governance instrument.[1396]

An important element that must be considered in the development of codes of conduct and, in general, co-regulatory frameworks, is the risk that the private sector, and especially large and powerful companies, take advantage of their position in order to produced biased codes that fail to protect third parties and lack effective enforcement mechanisms.[1397] It is thus necessary that, without disregarding the usefulness of codes of conduct, these mechanisms are completed or complemented by implementing appropriate enforcement mechanisms and that they are analysed by experts before they are approved in order to ensure that they are not just a mechanism to apparently ensure firms' compliance with rules and regulations while actually enabling the evasion of rules by private sector companies.[1398]

The fact that article 41 of the GDPR specifically refers to monitoring of private sector codes of conduct conveys how the Regulation aims to establish tools that ensure a system of effective accountability. However, considering that monitoring can also be delegated to private bodies, questions regarding the effectiveness of this system of accountability of controllers and processors arise.

All in all, if codes of conduct are to be used in order to ensure algorithmic accountability it is important to structure a system that guarantees their effectiveness. It is therefore necessary to include the participation of industry representatives but also civil society organisations and external experts when drawing up these codes in order to ensure that they are not only designed to safeguard firms' interests and serve as instruments to legitimise their actions.[1399] Additionally, oversight mechanisms must also be developed for the implementation and enforcement of said codes of conduct, probably including some degree of public sector participation. Said oversight mechanisms may, for instance include, public certification of codes of conduct.[1400]

[1396] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b.
[1397] BAMBERGER, K. A., "Regulation as delegation: private firms, decisionmaking, and accountability in the administrative state", *Duke Law Journal*, vol. 56, No. 2, 2006, pp. 428-429.
[1398] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1566.
[1399] KAMINSKI, M., "Binary governance…", *cit.*, 2019b, p. 1566; VEDDER, A. & NAUTS, L., "Accountability for the use of algorithms in a big data environment", *International Review of Law, Computers & Technology*, vol. 31, No. 2, p. 219.
[1400] KAMINSKI, M., "Binary governance…", *cit.*, 2019b, p. 1529.

4.4.1.3. Technical and organisational standards

Broadly speaking, standardisation[1401] aims to establish a series of rules in order to simplify activities, for example, the way in which the production of a certain item is carried out.[1402] The objective of standardising activities or products does not only make their development more efficient but also, and with regard to the elements that concern this research, provides a better framework for their control.[1403]

Said standards are generally referred to elements regarding the way in which algorithms are created and deployed, establishing certain requirements or mechanisms that must be complied with, such as the processes that must be set up in order to ensure personal data is correctly protected or that systems are structured in order to prevent and detect biases. Standards can be comprised within a framework of self-regulation, co-regulation or regulation.

There are already several examples of self-regulation for the prevention of algorithmic discrimination through standards. For example, IEEE's P7003 standard, which is still under development,[1404] offers a series of methodologies that should be implemented in order to detect and fight algorithmic bias when said bias affects especially vulnerable groups in a negative manner. Amongst the mechanisms offered are the following:

> "…benchmarking procedures and criteria for the selection of validation data sets for bias quality control; guidelines on establishing and communicating the application boundaries for which the algorithm has been designed and validated to guard against unintended consequences arising from out-of-bound application of algorithms; suggestions for user expectation management to mitigate bias due to incorrect interpretation of systems outputs by users (e.g. correlation vs. causation)".[1405]

[1401] The International Organization for Standardisation (ISO) defines a standard as: "*…documents that provide requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose*". See ISO, "Standards". Available on 3rd October 2019 at: https://www.iso.org/

[1402] ÁLVAREZ GARCÍA, V., "Introducción a los problemas jurídicos de la normalización industrial: normalización industrial y sistema de fuentes (1)", *Revista de Administración Pública*, No. 147, 1998, p. 309.

[1403] BARON, J. & SPULBER, D. F., "Technology standards and standard setting organizations: introduction to the Searle Center database", *Journal of Economics & Management Strategy*, No. 27, 2018, pp. 463-464.

[1404] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, p. 42.

[1405] IEEE, "IEEE P7003 standards for algorithmic bias considerations", *IEEE*. Available on 2nd October 2019 at: https://standards.ieee.org/

Standards generally refer to the way in which production is organised and therefore would mostly apply to the algorithm creation and deployment processes. However, since one of the issues that has been detected to be quite relevant in generating algorithms that produce discriminatory results or that reinforce pre-existing structures of disadvantage is the lack of diversity in the tech workforce[1406] it would be convenient for standards organisations to develop rules towards increasing diversity in the tech workforce and train employees on the ethical and social implications of algorithms. The increasing social demand for diversity has led many tech companies to develop programmes aimed towards hiring a more inclusive workforce.[1407] Hence, it is important to consider the possible incentives that standards organisations may have in establishing these types of rules. In any case, public-private partnerships with said organisations should also be developed in order to promote and ensure that diversity-related standards are set out.

Technical and organisational standards can complement codes of conduct for they are more precise and therefore easily actionable than the general principles that appear in codes of conduct. Once again however, the extent to which setting up standards will actually be effective in order to prevent and deal with situations of algorithmic discrimination will largely depend on the number and degree to which firms adhere to said standards.

While existing standards have been developed by private organisations, it is very likely that European institutions will also set requirements for the development of technical and organisational standards in order to prevent, amongst other risks, algorithmic discrimination.[1408] Said requirements can be carried out through the procedure set by article 10 of the 1025/2012 Regulation on European Standardisation[1409] which allows the EU

[1406] NOBLE, S. U., *Algorithms of Oppression…*, *cit*., 2018, p. 2.
[1407] APPLE, "Inclusion and diversity", *cit.*, 2017; WILLIAMS, M., "Facebook 2018 diversity report…", *cit.*, 2018; GOOGLE, "Google diversity annual report 2018", 2018; MCINTYRE, L., "Diversity and inclusion update…", *cit.*, 2018.
[1408] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1601; KAMARA, I., "Co-regulation in EU personal data protection: the case of technical standards and the privacy by design standardisation 'mandate'", *European Journal of Law and Technology*, vol. 8, No. 1, 2017, pp. 6-7.
[1409] Regulation (EU) 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation, amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council.

Commission to request European standardisation organisations to draft European standards.[1410]

The GDPR also offers a regulatory backdrop for technical and organisational standards to be developed by the European Commission.[1411] There are also several provisions throughout the regulation that actually set standards although they are not explicitly categorised as such.[1412] For instance, article 25 requires "data protection by design and by default", which must be implemented while also taking into consideration the principles contained in article 5.[1413] The data protection by design and by default mandate requires controllers to integrate data protection principles, such as data minimisation and purpose limitation, and introduce privacy safeguards in systems' initial design. Another GDPR provision that has indirectly become a technical standard is article 9, seeing as not only can you not use special categories if data as input variables but the system must be designed not to use them. In this regard, the notion of "equality by design", which was addressed in part I should also be developed as a standard.

4.4.2. Control mechanisms

4.4.2.1. Certification mechanisms

i) *General issues*

Compliance with standards is generally controlled and verified through certification systems. EDWARDS and VEALE have suggested two complementary ways of carrying out algorithmic certification:

- "Certification of the algorithm as a software object by (a) directly specifying either its design specifications or the process of its design, such as the expertise involved (technology-based standards) and/or (b) specifying output-related requirements that can be monitored and evaluated (performance-based standards); and

[1410] Article 10 of the Regulation on European standardisation: "The Commission may within the limitations of the competences laid down in the Treaties, request one or several European standardisation organisations to draft a European standard or European standardisation deliverable within a set deadline. European standards and European standardisation deliverables shall be market-driven, take into account the public interest as well as the policy objectives clearly stated in the Commission's request and based on consensus. The Commission shall determine the requirements as to the content to be met by the requested document and a deadline for its adoption".
[1411] Article 43, paragraph 9 of the GDPR.
[1412] KAMARA, I., "Co-regulation in EU personal data protection…", *cit*., 2017, p. 8.
[1413] *Ibidem.*

- Certification of the whole person or process using the system (system controller) to make decisions, which would consider algorithms as situated in the context of their use."[1414]

Certification was traditionally carried out by public administrations but is currently in many cases developed by private organisations[1415] that can be authorised by public entities or that operate in a completely private manner.

A common practice amongst certification bodies is to provide a trust or quality seal to those products or firms that are deemed to comply with a series of requirements after being audited. However, this form of private control has been proven to be quite ineffective in the past seeing as the money certification bodies obtain, generally comes from the fees that are paid by their members. It is therefore not convenient for certification bodies to establish very demanding requirements in order to obtain the seal or harsh sanctions when seal rules are broken.[1416]

However, there are also certain organisations that operate in a more independent manner and do not have any apparent incentives to be especially lenient with the products and firms that undergo their certification processes. Some of these organisations are non-profit while others simply ensure that they do not depend financially on those companies whose products are subjected to their certification processes.

In the context of algorithmic oversight, there are several organisations that generally police algorithmic fairness but that do not specifically focus on certification. Any individual that detects or suspects an algorithm to be unfair can bring said case to the attention of these organisations for them to examine but they do not offer certification mechanisms for firms that create or use data processing technologies. This is the way in which, for instance, Algorithm Watch[1417] and Algorithmic Justice League[1418] work.

[1414] EDWARDS, L. & VEALE, M., "Enslaving the algorithm: from a 'right to an explanation' to a 'right to better decisions'?", *IEEE Security & Privacy*, vol. 16, No. 3, 2018, p. 52.
[1415] DARNACULLETA I GARDELLA, M., *Autorregulación y Derecho Público…*, *cit.*, 2005, pp. 138-139.
[1416] EDWARDS, L. & VEALE, M., "Enslaving the algorithm…", *cit.*, 2018, p. 52.
[1417] ALGORITHM WATCH. Available on 17th October 2019 at: https://algorithmwatch.org/en/
[1418] ALGORITHMIC JUSTICE LEAGUE. Available on 17th October 2019 at: https://www.ajlunited.org/

However, certain firms specifically dedicated to certifying algorithms are also appearing. For instance, O'Neil Risk Consulting & Algorithmic Auditing (ORCAA)[1419] specifically focuses on offering auditing and certification services for algorithms and has developed its own quality seal.

While these organisations might lead the path towards a system of truly independent certification that would undoubtedly prove an essential mechanism for algorithmic oversight, there are still many elements that could present important barriers for the development of a really effective system of algorithmic certification. These mechanisms are voluntary and the development and use of algorithms that affect humans is increasingly disperse and cannot easily be narrowed down to a specific sector or set of firms. These factors are combined with the costs that subjecting software products for certification might entail for firms and act as deterrents for the development of a solid and effective algorithmic certification framework.

With regard to the costs that can be caused by certifying algorithms, it is important to highlight that a crucial element in hindering the development of algorithmic governance through voluntary certification is the rapid development of products in the software industry and the costs that are therefore generated by having to wait a certain period of time until the certification process is finalised. These costs are generally assumed when the products subjected to certification systems are built for sectors that are more heavily regulated and that entail greater risks[1420] and in which software development periods are generally longer than for software aimed towards final consumer use, which means that firms have longer to prepare for and run certification mechanisms.[1421] However, a very different reality appears regarding the certification of software products that are aimed to be distributed to final users. In this case, FERREIRA found that the amount of time that is required for a software system to undergo certification has in some cases even led to the release of later versions of the software than the one that was actually certified.[1422]

[1419] O'NEIL RISK CONSULTING AND ALGORITHMIC AUDITING (ORCAA), "Services", *cit.*, 2019.
[1420] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, p. 43.
[1421] FERREIRA, G., "Software certification in practice: how are standards being applied?", paper presented at the *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*, Buenos Aires, 2017, pp. 100-101.
[1422] *Idem*, p. 101.

In addition, with regard to certifying "the algorithm as a software object",[1423] it is important to highlight that the costs regarding certification mechanisms that must be considered are not just those associated to the waiting times that subjecting a product to certification might entail but also the actual cost of developing algorithm certification systems. The development of certification mechanisms can be very expensive, especially if they manage to get close or even achieve a zero-error margin.[1424]

Moreover, another difficulty that generally applies to the development of certification systems is the fact that it is very difficult to develop mechanisms that control of algorithmic discrimination. While it is possible to develop mechanisms that guarantee that the algorithm is doing exactly what it has been instructed to do, it is not as simple to certify for algorithmic fairness and non-discrimination.[1425] JOOSTEN differentiates between functional and qualitative control over algorithms and indicates that it is very difficult to develop the latter for it is necessary to, for example, determine what constitutes discrimination, which in many cases does not only require for legal instruments to specify what constitutes discrimination but also, in many cases, to interpret said provisions and determine whether a particular case is comprised under a prohibition to discriminate or whether differentiating between members and non-members of disadvantaged group is justified.[1426]

A recurring element that arises with regard to the certification of software systems is the risk that, by opening the black box, proprietary information will be released. It is therefore essential to develop systems that are able to certify that an algorithm is non-discriminatory without making the source code public. Several proposals have already been put forward by the literature towards this end.[1427] Nonetheless, particularly when algorithms can result in direct negative effects for individuals, it might be justifiable to use certification mechanisms that require full algorithmic transparency.

Finally, certification processes should also require processors and controllers to comply with certain organisational requirements, namely a diverse tech force. In this sense, it could be

[1423] *Ibidem.*
[1424] JOOSTEN, J., "Control funcional y control cualitativo de los algoritmos en la administración pública", paper presented at the *II Seminario internacional DAIA*, 10th and 11th October 2019. Slides available on 6th July 2020 at: https://www.phil.uu.nl/
[1425] JOOSTEN, J., "Control funcional y control cualitativo de los algoritmos en la administración pública", *cit.*, 2019; KROLL, J. A. *et al.*, "Accountable algorithms", *cit.*, 2017, p. 679.
[1426] KROLL, J. A. *et al.*, "Accountable algorithms", *cit.*, 2017, p. 679: "…for the tools to show that such systems meet policy goals, policymakers must determine the substantive properties that the systems should have…"
[1427] *Idem*, p. 705.

possible to require firms to comply with a certain degree of diversity along different axes in order to be certified as a company that respects equality in their organisation.

ii) *Certification in the GDPR*

Article 42, paragraph 1 of the GDPR indicates that "member states, the supervisory authorities, the Board and the Commission shall encourage, in particular at Union level, the establishment of data protection certification mechanisms and of data protection seals and marks". In addition, paragraph 3 explicitly states that certification mechanisms are voluntary. Thus, as with codes of conduct, the introduction of certification mechanisms is not mandatory. Therefore, considering the very elevated costs that come associated to the development of certification mechanisms and, in particular, the difficulties of developing systems that certify qualitative elements such as the principle of data minimisation or that an algorithm respects the right to equality and non-discrimination, it is possible that when algorithms are used in ways which generate a especially high risk of harming human rights, private firms choose to not subject said algorithms to certification mechanisms.

Nonetheless, the provisions contained in articles 42 and 43 of the GDPR do set provisions which ensure that, when firms choose to certify their software products or be certified as a whole, the certification process is trustworthy and effective. In this sense, the GDPR establishes limitations regarding the type of organisations that can carry out certification processes. Hence, only Data Protection Authorities or accredited bodies can certify that a certain data processing technique, a controller or a processor complies with the provisions of the GDPR. A particularly relevant point is the possibility that private bodies which are accredited carry out certification procedures. Requiring that public institutions accredit private certification bodies introduces a higher degree of control over private certification bodies and thus offers a greater guarantee of the effectiveness of certification processes.

General trust in certification processes will also be achieved to a greater extent if the certifying body is either accredited by a public institution or is a public institution.[1428] Conflicts of interest may arise when Data Protection Authorities carry out certification schemes since they would be in charge of both certifying data processing systems and of

[1428] RODRIGUES, R. *et al.*, "The future of privacy certification in Europe: an exploration of options under article 42 of the GDPR", *International Review of Law, Computers & Technology*, vol. 30, No. 3, 2016, p. 261, 263.

regulating and imposing sanctions on the same firms whose processes they have certified.[1429] Nonetheless, certifying a certain algorithm or firm does not mean that it is possible to control for all the possible outcomes in automated processing. Additionally, if certification and penalty-imposing activities are assigned to different departments the conflict would also be significantly reduced in any case.

All in all, implementation of the certification mechanisms contained in the GDPR will help to detect and prevent algorithmic discrimination to the extent that the provisions in the GDPR can serve to said purpose. Since the purpose of certification, as it is regulated in the GDPR, is to demonstrate compliance with the Regulation, only to the extent that, for example, the prohibitions of articles 9 and 22; the individual rights analysed in the previous chapter and the general principles of article 5 help to prevent and deal with instances of discrimination, will these certification mechanisms also contribute to the algorithmic anti-discrimination framework.

4.4.2.2. Data protection impact assessments

Impact assessments have been introduced as mandatory requirements for the development of many different activities. The context within which this mechanism has become more relevant is probably environmental protection.[1430] However there are also many other areas in which this regulatory instrument has been introduced. For example, gender equality impact assessments are required in Spain to be developed alongside all legislative proposals.[1431]

The International Association for Impact Assessment defines this mechanism as "the process of identifying the future consequences of a current or proposed action."[1432] Drawing from this very general definition, impact assessments can therefore be framed within self-regulatory, co-regulatory or state intervention forms of governance. Within the GDPR, articles 35 and 36 are dedicated to the regulation of Data Protection Impact Assessments (DPIAs), which, while

[1429] *Idem*, p. 262: "Their potential involvement in directly awarding data protection seals to data controllers (presumably, of a fee, even if nominal) risks a conflict of interest with regard to their mission, because they will be at the same time regulators and certifiers (a 'function creep' effect). Difficulties might arise in cases where a DPA might have to penalize a data controller, certified by it, for its personal data processing. It is therefore important for DPAs to safeguard their role in the EU data protection system, a task that could be compromised by their simultaneous role as data protection certifiers".

[1430] TONER, H., "Impact assessments and fundamental rights protection in EU law", *European Law Review*, No. 3, 2006, p. 316.

[1431] Article 19 of Organic Act 3/2007 for the effective equality of women and men.

[1432] INTERNATIONAL ASSOCIATION FOR IMPACT ASSESSMENT, "IAIA: Leading the global network for impact assessment", 2019. Available on 28th October 2019 at: https://www.iaia.org

still framed within a co-regulatory governance scheme, differ substantially from other mechanisms, such as certification, in that DPIAs are mandatory in certain cases. The regulation of DPIAs contained in articles 27 and 28 of the Directive for data protection in law enforcement is very similar to the GDPR. Consequently, while the following paragraphs only focus on the latter, the analysis carried out is also applicable to the Directive.

Article 35.1 GDPR states the following:

> "Where a type of processing […] is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data."

The main problem regarding article 35 is interpreting what constitutes a "high risk to the rights and freedoms of natural persons". Hence, in order to determine in which cases DPIAs are mandatory, that is, processing that generates a high risk for rights and freedoms of individuals, the A29WP refers and draws from the specific cases established by article 35.3:

> "(a) systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person;
>
> (b) processing on a large scale of special categories of data referred to in article 9(1), or of personal data relating to criminal convictions and offences referred to in article 10;
>
> (c) a systematic monitoring of a publicly accessible area on a large scale."

EDWARDS and VEALE conclude that this definition of high risk for the rights and freedoms of individuals will likely encompass most machine learning systems.[1433]

The way in which DPIAs are regulated clearly shows how the GDPR is aimed towards protecting other fundamental rights and, in particular, the right to equality and non-discrimination through the protection of personal data as an expression of privacy. In this sense, the A29WP indicated that, although DPIAs as regulated in the GDPR are mainly

[1433] EDWARDS, L. & VEALE, M., "Slave to the algorithm?…", *cit.*, 2017, pp. 77-78.

aimed towards preventing harms on the rights to privacy and data protection, the protection of other fundamental rights such as non-discrimination is also amongst the objectives of this mechanism.[1434] This is clearly portrayed in the above transcribed paragraphs (a) and (b) of article 35.3.

Once the DPIA has been carried out, controllers will have to consult the relevant supervisory authority in the member state when the Impact Assessment determines that "processing would result in a high risk in the absence of measures taken by the controller to mitigate the risk."[1435] If said supervisory authority considers that the that the processing is contrary to the GDPR it may either establish certain requirements that the controller should implement to diminish the risk or even "impose a temporary or definitive limitation including a ban on processing".[1436]

One of the most significant shortcomings of the DPIA system established by the GDPR is the fact that these impact assessments do not have to be released to the public. While the Regulation does encourage release, it does not contemplate it as a mandatory measure and only considers the possibility of publishing certain parts or a summary of the impact assessment.[1437] This particular form of lack of transparency hinders the possibility that third parties will be able to enact oversight mechanisms and help data protection authorities in preventing that particularly risky forms of processing are carried out.[1438]

4.4.2.3. Re-certification, DPIA reviews and audits

Given the fact that machine learning systems are constantly self-developing, it is important not only to have certain mechanisms or tools that assess their compliance with applicable regulations prior to deployment but also after they have been implemented for some time. While the GDPR does foresee these *ex post* control tools it fails to provide a comprehensive and accountability framework.

[1434] ARTICLE 29 WORKING PARTY, "Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679", 17/EN, WP 248 rev.01, 4th October 2017, p. 6.
[1435] Article 36.1 GDPR.
[1436] Article 36.2 GDPR.
[1437] ARTICLE 29 WORKING PARTY, "Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679", *cit.*, 2017, p. 18.
[1438] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1608.

Certification is, at least in theory, the most effective mechanism for controlling that algorithms still comply with the regulatory mandates applicable to them after they have been deployed. This is due to the fact that certification is only valid for a maximum period of three years before it has to be renewed.[1439] This provision ensures that there is a periodic control of compliance with the GDPR. However, considering that certification is voluntary under the GDPR, the actual effectiveness of certification renewal will depend on the enforceability of the collaborative governance system, that is, the extent to which firms and the general public feel that the provisions of the GDPR, even when voluntary, should be binding and complied with.

The Regulation also includes mandatory DPIA reviews "when there is a change of the risk represented by processing operations",[1440] amongst which the A29WP includes changes in the technology used or in the purpose for which the data is being processed.[1441] Once again, the vagueness of this provision and lack of enforcement and control mechanisms hinders general compliance and, hence, its effectiveness.

Audits are also mentioned in several provisions throughout the GDPR[1442] and are mostly referred to as internal tools for compliance control that are to be carried out by data controllers with the aid of data processors. The A29WP guidelines on automated processing also indicate algorithmic auditing (carried out both by controllers and processors as well as third parties) as an appropriate safeguard that should be considered when carrying out decisions based solely on automated processing, including profiling.[1443] In addition, data protection audits carried out by supervisory authorities are also foreseen but with no indication of content or periodicity.

Finally, and specifically with regard to controlling that the algorithms used do not discriminate against individuals, the A29WP guidelines on automated processing also establish as an appropriate safeguard for instances of decisions based on solely automated processing, including profiling, that controllers should implement "regular quality assurance

[1439] Article 42.7 GDPR.
[1440] Article 36.11 GDPR.
[1441] ARTICLE 29 WORKING PARTY, "Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679", *cit.*, 2017, pp. 13-14.
[1442] Articles 28.1(h), 39.1(b), 47.2(j) and 58.1(b) GDPR.
[1443] ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", *cit.*, 2018, p. 32.

checks of their systems to make sure that individuals are being treated fairly and not discriminated against, whether on the basis of special categories of personal data or otherwise."[1444]

4.4.3. Enforceability mechanisms

4.4.3.1. Ethics committees and data protection officers

While ethics committees are not specifically mentioned in the GDPR, they are worth mentioning as they are becoming increasingly common, particularly in big tech companies, and are precisely aimed towards controlling that automated systems respect the fundamental rights and freedoms of individuals. Ethics committees are generally a form of self-regulation that can be set up within a firm or within a certain sector. They can be useful in order to control for compliance with codes of conduct and standards. In many cases, firms have established said committees in order to deal with scandals regarding risks for human rights generated by the use of algorithms,[1445] an element that unavoidably puts into question the real effectiveness of said mechanism of control as it draws concerns that their main objective is to circumvent government regulation and exert damage control when public relations crises take place.[1446] Moreover, the ethics boards that have been set up during the past few years as responses to the public release of information on controversial uses of data, are still yet to prove their impact as more than mere public image protection mechanisms.[1447]

The GDPR does however set up a specific body (data protection officer) within both controllers and processors to control for compliance with European Union and member state data protection rules as well as with other informal regulation mechanisms, including firms' internal rules, standards and codes.[1448] Designation of a data protection officer is always compulsory for public authorities (except courts) and for private organisations when they develop particularly risky forms of processing.[1449] The objective of creating the role of data

[1444] *Ibidem.*
[1445] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, p. 43.
[1446] WAGNER, B., "Ethics as an escape from regulation: from ethics-washing to ethics-shopping?", in HILDEBRANDT, M., (ed.), *Being profiled. Cogitas Ergo Sum*, Amsterdam, Amsterdam University Press, 2018, pp. 84-85.
[1447] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, p. 43.
[1448] Articles 37-39 GDPR; KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, p. 1596.
[1449] Article 37.1 GDPR: "The controller and the processor shall designate a data protection officer in any case where: the processing is carried out by a public authority or body, except for courts acting in their judicial

protection officers within organisations is to establish an internal authority that will act as a connection between data protection agencies and data controllers and processors in the hope and with the objective of enhancing compliance with the GDPR through this form of collaborative governance.

4.4.3.2. The European Data Protection Board and Data Protection Authorities

In line with the general tendency to create independent public authorities to control that both public and private organisations comply with certain regulatory aspects, such as competition, the GDPR establishes the European Data Protection Board (EDPB), which substitutes the Article 29 Working Party. The role of the EDPB, which is composed by the European Data Protection Supervisor[1450] and the heads of each member state's supervisory authority,[1451] is to ensure general compliance with the Regulation and issue Guidelines and other documents and generally advise on the interpretation of the GDPR's provisions with the aim of helping to harmonise data protection rules in the different European member states.[1452] In this sense, one of its main objectives is to foster cooperation between member states' Data Protection Authorities. More specifically, with regard to the mechanisms that have been analysed in this section, the EDPB also encourages drawing up and implementing of codes of conduct and certification mechanisms and controls them.

Data Protection Authorities are set up in each member state with the objective of controlling compliance with the GDPR and internal data protection rules within each country. Amongst other tasks, they control certification systems, intervene in DPIAs and can even limit or ban certain data processing activities and impose administrative fines.[1453] On paper, these institutions have a very wide range of possibilities in order to control that public and private organisations comply with the GDPR and even to ensure that no discriminatory processing

---

capacity; the core activities of the controller or the processor consist of processing operations which, by virtue of their nature, their scope and/or their purposes, require regular and systematic monitoring of data subjects on a large scale; the core activities of the controller or the processor consist of processing on a large scale of special categories of data pursuant to Article 9 and personal data relating to criminal convictions and offences referred to in Article 10."

[1450] The European Data Protection Supervisor is established by Regulation 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC and is "responsible for ensuring that the fundamental rights and freedoms of natural persons, and in particular their right to data protection, are respected by Union institutions and bodies" (Article 52.3 of Regulation 2018/1725).

[1451] Article 68.3 GDPR.

[1452] Article 70 GDPR.

[1453] Article 58 GDPR.

takes place, however, and while they do carry out significant actions when complaints are lodged before them, their oversight capabilities still fall short from being able to address and prevent the risks generated by the automated or semi-automated processing of personal data.

4.4.3.3. Penalties

Penalties are necessary both in order to enact the right to an effective remedy and to make governance systems effective. The Directive on data protection in law enforcement recognises the right to compensation in article 56 and, in article 57, establishes a general mandate for EU member states to establish systems of penalties that should be "effective, proportionate and dissuasive".

The GDPR, which regulates these aspects in articles 82, 83 and 84, logically establishes a more comprehensive and detailed system of compensation and administrative fines, which is approached as a tiered system both, depending on the gravity of the infringement and the size of the undertaking. Administrative fines can be of up to €20,000,000 or 4% of the undertaking's total annual turnover. As of July 2020, 325 penalties have been imposed, which have ranged from €118 for the disclosure of a third party's personal data without sufficient legal basis,[1454] to the €204,600,000 (still provisional) fine that the UK's Information Commissioner's Officer has announced it intends to fine British Airways with for failing to implement sufficient technical and organisational standards to ensure information security, which led to a data breach.[1455]

[1454] GDPR Enforcement Tracker. Available on 6th July 2020 at: https://www.enforcementtracker.com

[1455] INFORMATION COMMISSIONER'S OFFICE, "Intention to fine British Airways £183.39m under GDPR for data breach", 8th July 2018. Available on 6th July 2020 at: https://ico.org.uk/: "The proposed fine relates to a cyber incident notified to the ICO by British Airways in September 2018. This incident in part involved user traffic to the British Airways website being diverted to a fraudulent site. Through this false site, customer details were harvested by the attackers. Personal data of approximately 500,000 customers were compromised in this incident, which is believed to have begun in June 2018".

CHAPTER III. THE PRIVACY FRAMEWORK: SHORTCOMINGS AND TENSIONS

The European regulatory framework on data protection offers a series of mechanisms that can be useful for controlling and preventing some of the risks generated by the automated processing of personal data and the decisions and profiles that may result from it. However, the GDPR and Directive on data protection in law enforcement fall short in establishing an effective system of individual rights and algorithmic accountability. Especially when it comes to the risks to other fundamental rights, especially the rights to equality and non-discrimination, the data protection framework fails to offer sufficient safeguards, accountability and remedies, mainly because it is not articulated as a non-discrimination instrument, but also because many of its provisions have, in practice, turned out to be soft-law instruments.

The following pages analyse the different shortcomings of the privacy framework, in general, but also, specifically, as a mechanism to protect against the risks that algorithmic processing and decision-making generates for the rights to equality and non-discrimination. The first part of this chapter thus focuses on the general shortcomings of these instruments; the second part analyses the insufficiencies generated by an over-reliance on the informational self-determination, also known as individual rights, approach to personal data protection; the third part establishes the reasons why data protection approaches, as they are currently structured, are insufficient to apply to the public use of algorithms and the fourth part briefly highlights the ways in which the governance mechanisms contained in data protection instruments fail to effectively establish algorithmic accountability mandates. The fifth part examines the tensions between the way in which privacy protections are currently articulated and the regulatory needs that an effective protection of the rights to equality and non-discrimination entails. This chapter ends with a brief reference to the way in which, given the current regulatory framework, combining the tools offered by personal data protection and equality and non-discrimination protection instruments could be useful in preventing and fighting instances of algorithmic discrimination.

1. GENERAL SHORTCOMINGS OF THE PRIVACY APPROACH

1.1. THE UNREALISTIC EXPECTATIONS OF ANONYMISATION

Data protection rules regarding the prohibition to process certain types of information do not apply when a dataset is anonymised and thus the individuals whose data is being processed cannot be identified.[1456] Many privacy regulations are largely framed from the perspective of considering that, once the dataset has been anonymised individuals are no longer at risk of being discriminated or suffering other harms.[1457] This approach clearly falls short of the needs brought by the new reality resulting from the development of new data processing technologies given the fact that they allow for the identification of individuals even after their data has been anonymised.[1458]

The first section in the previous chapter covered the issues that arise from the restrictive interpretation of what is considered an anonymised dataset within the GDPR. As it was stated in said chapter, the problem with anonymised datasets is not as much the possibility of extracting personally identifiable information from them but the way in which the data is processed and whether the collector or processor aim to purposefully identify individuals from the data provided in anonymised samples.

This is due to the fact that full anonymisation is impossible[1459] and that the risks of re-identification are constantly increasing as technology develops.[1460] Consequently, if informational privacy regulations continue to leave anonymised datasets out of their scope of application there are two likely negative outcomes. Firstly, the requirements for a dataset to be considered anonymised will become increasingly restrictive, thereby hampering the development of research that could be socially beneficial,[1461] especially when said research is developed by actors with less resources for whom achieving anonymisation standards is more costly. Secondly, a consequence of said restrictive interpretation is that most datasets fall under the scope of privacy regulations, meaning that, in order to be able to process data,

[1456] See, for example, Recital 26 of the GDPR.
[1457] YOUNG, E., "Educational privacy in the online classroom: FERPA, MOOCs, and the Big Data Conundrum", *Harvard Journal of Law & Technology*, vol. 28, No. 2, 2015, p. 553.
[1458] GIL GONZÁLEZ, E., *Big data, Privacidad y Protección de Datos*, *cit.*, p. 83.
[1459] ZIBUSCHKA, J. *et al.*, "Anonymization is dead – long live privacy", in ROßNAGEL, H., WAGNER, S. & HÜHNLEIN, D., (eds.), *Open Identity Summit 2019*, Bonn, Gesellschaft für Informatik, 2019, p. 72.
[1460] OHM, P., "Broken promises of privacy…", *cit*., 2010, p. 1776.
[1461] EL EMAM, K. & ÁLVAREZ, C., "A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques", *cit.*, 2015, p. 75.

powerful sectors will try to pressure governments to introduce exceptions to data processing prohibitions and restrictions that apply to them and to provide extensive interpretations to purpose limitation and data minimisation principles.[1462]

1.2. THE LIMITS OF PERSONAL DATA PROTECTION

1.2.1. Group profiling

Group profiling is closely associated to the lack of protection offered to anonymised data. Group profiles are developed with regard to individuals who have certain characteristics in common. Personal data protection regulations do not apply as long as group profiles do not include identifiable personal information. [1463] The amount of information that is currently available means that profiles are in many cases not constructed using personal data but by identifying certain traits or actions that characterise individuals as members of a group.[1464] Group profiles, which are constructed without using personal data, are then applied to specific individuals and can cause significant harms, especially, to members of disadvantaged groups.

1.2.2. Output data

While the GDPR focuses on protecting the input data that is processed, it fails to specifically address algorithmic output data, that is, the inferences made by the system, as a type of personal data.[1465] It seems fairly obvious that the definition of personal data provided by the GDPR[1466] does include the resulting inferences through which are or can be linked to the individual. However, the current system does not offer sufficient mechanisms to detect and challenge inaccurate or generally harmful inferences.[1467] Additionally, there are no rules determining the way in which output data should be evaluated and no mandatory safeguards that controllers and processors should implement on output data.

[1462] *Ibidem*.
[1463] SCHREURS, W. *et al.*, "Cogitas, ergo sum…", *cit.*, 2008, p. 243.
[1464] KOOPS, B. J., "The problem with European data protection law", *International Data Privacy Law*, vol. 4, No. 4, 2014, p. 257: "The key feature of profiles is that they do not necessarily relate to individuals, but often to groups ('someone with characteristics x, y, and z'), which makes them non-personal data (until they are applied to identified individuals) and hence the creation and much of the processing of profiles traditionally falls outside of data protection law."
[1465] WACHTER, S. & MITTELSTADT, B. D., "A right to reasonable inferences…", *cit.*, 2019, p. 499.
[1466] Article 4.1 GDPR: "Any information relating to an identified or identifiable natural person."
[1467] WACHTER, S. & MITTELSTADT, B. D., "A right to reasonable inferences…", *cit.*, 2019, p. 530.

1.2.3. Failure to focus on varieties of processing

The GDPR only focuses on the actual process when it prohibits solely automated-decision making. However, it generally approaches the protection of personal data in a very limited way and while regulating and establishing restrictions to the processing of personal data, it does not focus on the different types of processing that could take place.[1468] It therefore fails to determine whether certain types of modelling processes are too risky or inappropriate for certain purposes.

2. THE SHORTCOMINGS OF THE INFORMATIONAL-SELF DETERMINATION APPROACH

As the chapter on "technological due process"[1469] indicated, the data protection legal tradition has mostly focused on informational self-determination as the main mechanism through which to structure and regulate specific tools for the effectiveness of the right to data protection.[1470] This approach has unquestionable benefits such as the fact that it recognises individuals' agency, thereby providing them with the dignity and autonomy that automated tools can sometimes deny them.[1471]

In a more practical sense, these individual rights can provide individuals with the necessary tools to challenge erroneous, unfair or illegal automated or semi-automated decisions that affect them.[1472] Moreover, the exercise by individuals of their rights to a technological due process can have a domino effect whereby others also decide to exercise said rights hence disclosing certain data collection and processing practices[1473] and even leading to legislative changes[1474] and providing the necessary information for third parties to carry out system-

[1468] ŽLIOBAITĖ, I. & CUSTERS, B., "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models", *cit.*, 2016, p. 188: "Note that the law only addresses the data, not the modeling process. The only legal requirement relating to the process is that people should not be subjected to (entirely) automated decision-making."
[1469] CITRON, D. K., "Technological due process", *cit.*, 2008, pp. 1249-1313.
[1470] KOOPS, B. J., "The problem with European data protection law", *cit.*, 2014, p. 251.
[1471] KAMINSKI, M., "Binary governance…", *cit.*, 2019b, pp. 1553-1554.
[1472] *Idem*, pp. 1555.
[1473] KOOPS, B. J., "The problem with European data protection law", *cit.*, 2014, p. 252: "The fact that within two months after the Google Spain ruling, Google received 91,000 removal requests involving 328,000 URLs14 shows that data subjects are actively exercising their right to request erasure".
[1474] DIMITROVA, D. & DE HERT, P., "The right of access under the police directive…", *cit.*, 2018, p. 115: "…the disclosure of different illegalities related to the processing of one' s data could have a wider impact, i.e. trigger political, judicial and policy-making action by raising awareness about the processing operations which affect the public at large. An example is the case of Max Schrems' s access to his Facebook data which lead to more

wide oversight. However, as this section conveys, the individual rights perspective fails to provide the necessary mechanisms for a comprehensive and general protection against the risks generated by data processing technologies.

2.1. THE MYTH OF CONSENT AND THE PRIVACY PARADOX

Individual consent for data collection and processing has become the main way through which data protection regulatory instruments have specified the idea of informational self-determination.[1475] However, as the following paragraphs will show, this ideal of providing individuals with tools to manage their personal data by granting or denying consent whenever they see fit has proven to be little more than a myth.[1476]

Relying on informational self-determination is necessary to a certain extent for it recognises the individuality, autonomy and freedom of those people it applies to and provides them with mechanisms to protect their sphere of privacy. However, the European data protection framework has been built by placing individual action as its main mechanism of protection, which unavoidably leads to the existence of a series of very significant gaps that heavily limit the effectiveness of the EU data protection framework and largely prevent these instruments from being able to protect other rights that can also be endangered through data processing and automated decision-making. The following paragraphs convey how a series of market failures, such as irrational behaviours and asymmetric information, arise when individuals find themselves in contexts in which they have to or can share their personal data.

The ineffectiveness of consent as basis for lawfulness is for instance depicted in the phenomenon labelled as "the privacy paradox".[1477] The direct benefits that sharing personal information sometimes brings leads to a series of contradictions in individuals' behaviour that have been widely documented by the literature. When individuals are directly asked, they share concerns regarding privacy and the use of data processing technologies, however when their behaviour is observed it seems like these concerns are not as pressing as they express

---

Facebook users claiming access to their data and to judicial proceedings and legislative changes such as striking down the Safe Harbour and replacing it with the Privacy Shield."

[1475] KOOPS, B. J., "The problem with European data protection law", *cit.*, 2014, p. 251.

[1476] *Idem*, p. 252.

[1477] ATHEY, S., CATALINI, C. & TUCKER, C., "The digital privacy paradox: small money, small costs, small talk", *MIT Sloan Research Paper No. 5196-17*, 2017. Available on 21st May 2019 at: https://papers.ssrn.com/

them to be seeing as they will gladly share information even when offered very small incentives for it.[1478]

It is nevertheless important to highlight that the existence of this difference between expressed privacy preferences and actual behaviours does not mean that individuals do not care at all about privacy but that they face a trade-off between privacy and obtaining benefits such as accessing certain types of online content.[1479] YOUNG eloquently illustrates this conflict when it specifically applies to the benefits brought by the development of data processing technologies by indicating that we face "the dilemma between creepiness and innovation that big data has generated. We want privacy (or maybe just anonymity), but we also want the benefits of big data".[1480]

Moreover, several studies have shown that when individuals balance the benefits and costs in the trade-off between privacy and the benefits of sharing data, attitudes that show a certain degree of concern with regard to their personal information are conveyed seeing as, in general, they are willing to pay a prime for using services in websites that clearly display and offer better privacy policies.[1481]

The choices made by individuals are also heavily mediated by the fact that the possibility of denying consent for data collection and processing is no more than an illusion when providing consent is a prerequisite to access a certain service.[1482] This is probably one of the determining factors in the distortion between individuals' privacy preferences and attitudes in cases in which access to a service, such as the main social networks, while not absolutely necessary for an individual, has become a normal part of social interactions. Hence, the perceived effects of not providing consent are far worse than allowing organisations to process one's personal data. This is true, in particular, if we consider that the effects of data processing are generally not immediately nor directly experienced by individuals. In addition, in many cases, the lack of alternatives to the provision of some services means that there is in fact no real choice, particularly when these services are essential for an individual.

[1478] *Idem*, p. 4.
[1479] KOKOLAKIS, S., "Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon", *Computers & Security*, vol. 107, 2017, p. 128.
[1480] YOUNG, E., "Educational privacy in the online classroom…", *cit.*, 2015, p. 561.
[1481] KOKOLAKIS, S., "Privacy attitudes and privacy behaviour...", *cit.*, 2017, p. 125.
[1482] KOOPS, B. J., "The problem with European data protection law", *cit.*, 2014, p. 251.

2.2. ASYMMETRIC INFORMATION AND BURDENS

Informational self-determination frameworks are largely based on the idea of transparency. If individuals know which of their personal data is being processed and how, they will be able to challenge said processing and the decisions that result from it.[1483] However, as CRAWFORD and ANNANY indicate, "the imagined marketplaces of total transparency have what economists would call perfect information, rational decisionmaking capabilities, and fully consenting participants. This is a persistent fiction".[1484] The myth of consent is further enhanced by asymmetric information and the burdens that the informational self-determination framework poses on individuals.

In this sense, it is important to highlight that there are many cases in which individuals are not really aware of what it is they are agreeing to.[1485] This may result from the complex language that is sometimes used when explaining the terms of consent but also due to the difficulty of actually understanding how data is processed and used by organisations.[1486] In the end, providing consent does not necessarily entail it is meaningful or informed but that is has simply become a formality or step that individuals have to take in order to access a certain service.[1487]

Placing informational self-determination at the centre of data protection systems also results in a series of burdens for data subjects.[1488] While the existence of individual rights is a necessary mechanism for protection against the risks generated by the use of big data and its associated technologies, mainly focusing on this approach can render the whole system largely ineffective. The problems that understanding automated or semi-automated systems generates are not only crystallised through individuals' lack of awareness when consenting to the collection and processing of their personal data, but also through the difficulties that exercising their "technological due process"[1489] rights generate.

[1483] ANNANY, M. & CRAWFORD, K., "Seeing without knowing…", *cit.*, 2018, pp. 979-980.
[1484] *Idem*, p. 980.
[1485] KAMINSKI, M., "Binary governance...", *cit.*, 2019b, p. 1590: "One criticism of the FIPs (fair information practices) is that they can be substance-less, providing individuals the illusion of control while in practice allowing companies to do nearly anything as long as they have gotten individuals to click through an agreement."
[1486] KOOPS, B. J., "The problem with European data protection law", *cit.*, 2014, p. 252.
[1487] SCHREURS, W. *et al.*, "Cogitas, ergo sum…", *cit.*, 2008, p. 250.
[1488] EDWARDS, L. & VEALE, M., "Slave to the algorithm?…", *cit.*, 2017, p. 67.
[1489] CITRON, D. K., "Technological due process", *cit.*, 2008, pp. 1249-1313.

While there are cases that have proven the usefulness of individual rights for the control and protection of informational privacy, not all individuals are aware of the existence of these rights or know how to exercise them.[1490] The complicated nature of both algorithms and the individual rights framework designed to control data processing tools creates the perfect setting for data processors and controllers to avoid full compliance with the obligations triggered by the exercise of individual rights. If individuals are not fully aware of which of their personal data is held by controllers and processors and do not really understand the full extent of their rights, it will be easier for organisations using personal data to avoid fulfilling certain duties.

Finally, these models of protection based on the idea of individual empowerment[1491] can lead to increasing inequalities in the ability to protect informational privacy by members of different socioeconomic strata. SUNSTEIN argues that poorer segments of society are the most harmed by administrative paperwork.[1492] This idea can be extrapolated to the heavy burdens that are imposed on individuals regarding the exercise of rights that can allow the *ex ante* and *ex post* protection of their personal data and other harms caused by algorithmic processing. While wealthier individuals will have the resources and time to exercise their rights and protect their personal data, individuals with lower socioeconomic backgrounds will have greater difficulties in circumventing the data sharing requirements set up as quasi-mandatory or mandatory conditions for the access to certain services and to exercise *ex post* individual rights for the protection of their personal data.

2.3. CREATING SYSTEMIC INACCURACIES

The rights that the data protection framework provides individuals with can lead to the introduction of biases in algorithmic systems.[1493] For example, by exercising the right to rectification and erasure, individuals can alter certain accurate information on them with the objective of changing their profiles and gaming automated systems.[1494] If this practice becomes widespread it will lead to increasingly biased profiles.

[1490] KOOPS, B. J., "The problem with European data protection law", *cit.*, 2014, p. 252.
[1491] ANNANY, M. & CRAWFORD, K., "Seeing without knowing…", *cit.*, 2018, pp. 979-980.
[1492] SUNSTEIN, C., "Sludge and ordeals", *cit.*, 2018, p. 1859.
[1493] KAMINSKI, M., "Binary governance…", *cit.*, 2019b, p. 1581.
[1494] BAMBAUER, J. & ZARSKY, T., "The algorithm game", *cit.*, 2018, p. 35.

Moreover, even when individual rights are not purposefully used in order to game the system they can also lead to systemic inaccuracies that can, in turn, reinforce situations of inequality. If, as it was argued in the previous section, those who enjoy better economic circumstances are more likely to exercise data protection rights and, for example, select data processing opt out options, automated systems will contain more information on poorer individuals, thereby leading to an increased algorithmic control of these individuals' lives.[1495] Considering that socioeconomic elites will have better chances to opt out from being processed by especially risky algorithms, the differences between those who are able to exit the system and those who remain will probably reinforce pre-existing situations of inequality.

2.4. THE DIFFICULTY OF DETECTING SYSTEMIC ERRORS

Algorithmic discrimination results from biased or erroneous datasets, classifications that favour members of privileged population groups and sometimes even from accurate inferences that arise from social systems of structural discrimination. Whichever the origin of discriminatory outcomes, these types of biases are embedded in algorithms, which means that that this specific type of problem creates greater risks for groups (and for individuals as members of said groups) than for individuals independently considered.[1496] Hence, an individual rights approach, while appropriate to detect some specific cases of discrimination cannot be used as a system oversight framework.[1497] This does not mean that the detection of specific instances of discrimination cannot lead to spotting the domination narratives embedded in algorithms,[1498] however, since the individual rights framework is not designed with the idea of protecting and preventing the harms that data processing can cause on groups but specific individuals, it is much harder to detect and prevent discrimination as well as other systemic problems.[1499]

For example, the right to an explanation that the GDPR recognises individuals might provide them with a justification for the automated decision made but it will not necessarily convey information regarding the way in which the whole system works as it is not designed as an oversight mechanism. Furthermore, even if an individual explanation conveys more general knowledge on the way the system works, it might not necessarily provide sufficient

[1495] KAMINSKI, M., "Binary governance…", *cit.*, 2019b, p. 1581.
[1496] EDWARDS, L. & VEALE, M., "Slave to the algorithm?…", *cit.*, 2017, pp. 83-84.
[1497] KAMINSKI, M., "Binary governance…", *cit.*, 2019b, p. 1590.
[1498] *Ibidem*.
[1499] *Ibid*.

information to detect that a certain decision is based on a discriminatory rational. An individual explanation for a decision might justify said decision based on certain characteristics of an individual that will not directly include protected group membership. Hence, unless said individual is capable of spotting the correlation between the logic provided and protected group membership or acquires knowledge of other members of the same protected group that have also suffered negative consequences from that automated decision-making process, it will not be possible to identify the system as discriminatory.

Moreover, "technological due process"[1500] rights are generally only triggered after algorithms have been deployed and have already resulted in cases of discrimination. This is particularly problematic if we consider that one of the characteristics of machine learning algorithmic systems is their capacity to continuously change and develop.[1501] Hence, in order to control and fix systemic problems it is better to develop an effective oversight framework for algorithms that ensures good algorithm design and which controls algorithms on a regular basis.[1502]

Finally, there are cases in which individual rights are simply not operative. For example, when a system is designed to target poorer individuals with certain types of advertisements, it will be difficult to use individual privacy rights to detect that this is occurring, especially considering that there are no direct effects on individuals. It is very unlikely that these cases will be detected through the exercise of individual rights. Once the fact that low quality products and services are being targeted to members of vulnerable groups is detected, it will be possible to exercise, for instance, the right to restriction of processing, however, the difficulty of detecting these activities still remains. Moreover, even when detected, some instances of discrimination will doubtfully fall under unlawful forms of processing.[1503]

[1500] CITRON, D. K., "Technological due process", *cit.*, 2008, pp. 1249-1313.
[1501] ANNANY, M. & CRAWFORD, K., "Seeing without knowing…", *cit.*, 2018, p. 982.
[1502] KAMINSKI, M., "Binary governance…", *cit.*, 2019b, pp. 1558-1559; EDWARDS, L. & VEALE, M., "Slave to the algorithm?…", *cit.*, 2017, pp. 83-84.
[1503] WIEDEMANN, K., "Automated processing of personal data for the evaluation of personality traits: legal and ethical issues", *Max Planck Institute for Innovation and Competition Research Paper No. 18-04*, 2018, p. 25. Available on 29th July 29 2019 at: https://ssrn.com/: "The study found that (simulated) male users were shown more advertisements relating to high-paying jobs than female users when visiting websites associated with employment … Under EU anti-discrimination and data protection laws, the gender discrimination described in the study would probably not be unlawful. In particular, this is not a case of automated decision-making as prohibited in Art. 22(1) of the GDPR, since the mere display of discriminatory user-targeted contents does not produce immediate (legal or other) effects for the user and is, generally speaking, of a rather passive nature. It is unclear what role the anti-discrimination part of Recital 71 plays in this regard, since discrimination based on gender is not named explicitly in its wording."

3. PRIVACY APPROACHES ARE NOT APPROPRIATE FOR THE USE OF ALGORITHMS BY THE PUBLIC SECTOR

Existing informational privacy legal instruments are mainly designed as regulations to be applied to the private sector. This especially obvious in the case of the GDPR which, although applies to both the public and private sectors has failed to incorporate the requirements that are generally set for public bodies, especially when they exercise authority.[1504] In fact, the GDPR, following the trend set by the Data Protection Directive and other informational privacy European regulatory instruments, establishes a wide set of exemptions from data protection mandates whenever public duties are being exercised.[1505] For example, the prohibition to process special categories of data in a fully or semi-automated manner is lifted whenever said processing is carried out for "reasons of substantial public interest" (articles 9.2.g and 22.4 GDPR). Moreover, article 23 of the Regulation establishes a general clause according to which the rights included in articles 12 to 22 of the GDPR[1506] can be restricted in order to safeguard public security and interests.

While all the exemptions and right restrictions previously noted require respecting the data subject's rights and freedoms,[1507] they offer a wide spectrum of action within which public powers cannot only use algorithms as tools in decision-making processes but can also limit the disclosure of the information contained in said systems. This is especially striking if we consider that the processes that are currently being carried out by automated systems are analogous to traditional administrative processes to which public law applies a greater degree of requirements and guarantees which public administrations seem to be aiming to circumvent when using automated systems.

3.1. PRIVATE SECTOR LIMITS TO TRANSPARENCY FOR ALGORITHMS USED BY PUBLIC BODIES

In their paper "Algorithmic transparency for the smart city",[1508] BRAUNEIS and GOODMAN document a series of cases in which trade secrets and nondisclosure agreements were wielded

[1504] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, pp. 247-248.
[1505] SCHREURS, W. *et al*., "Cogitas, ergo sum…", *cit.*, 2008, p. 265.
[1506] The rights to information; access; rectification; erasure; restriction of processing; data portability; to object; and, not to be subjected to solely automated decision-making, including the subset of rights comprehended under said right-prohibition.
[1507] See, for example, arts. 22.2(b) and 23.1 GDPR.
[1508] BRAUNEIS, R. & GOODMAN, E. P., "Algorithmic transparency for the smart city", *cit.*, 2018, pp. 103-176.

by US public agencies as limits to the right to access public information, thereby justifying that algorithms used for public purposes were kept secret.[1509] A particularly striking element of their findings is the fact that public agencies seem to be willing to interpret the trade secret clauses applicable to public information in a much more extensive way than regulatory instruments and case law suggests they should be.[1510]

3.1.1. Intellectual property and the Spanish "energy social bond"

A paradigmatic example of the way in which public administrations do not respect traditional requirements and guarantees applicable to the public sector when using automated systems is the Spanish "energy social bond" case referred to in part I. As it was already explained in the first part of the dissertation, Spanish public authorities have developed an algorithm that automatically determines when an applicant meets the requirements set by the relevant regulatory instrument[1511] in order to be granted a discount rate in their energy bill.

The energy bond automated system produced erroneous results, as it rejected applicants who met said requirements. The mistakes resulted from the way in which the programme was set up, which deviated from the regulatory instrument it was supposed to be implementing. In other words, this system was effectively regulating and, not just that, but it was regulating *contra legem*.[1512] It is essential, at this point, to draw attention to the fact that said automated system is, in theory, designed to simply apply a legal rule to specific cases and can therefore be classified under the "automatic" systems category as they are not purely "autonomous",[1513] hence its complexity should be minimum.

Once these erroneous outcomes were detected, CIVIO, a Spanish civil society organisation, requested access to the algorithm's source code on the basis of the right of access to public information recognised in article 105.[1514] Article 12 of the Spanish Transparency Act[1515] establishes the right of access to public information. Article 13 establishes that the concept of "public information" comprises "any content or documents, whatever their format or support,

[1509] *Idem*, pp. 153-160.
[1510] *Idem*, pp. 156-157.
[1511] Royal Decree 897/2017, 6th October, which regulates the concept of vulnerable consumer, the social bond and other protection measures for domestic electricity consumers.
[1512] DE LA CUEVA, J., "El derecho a no ser gobernados mediante algoritmos secretos", *cit.*, 2019.
[1513] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, p. 228.
[1514] MESTRE DELGADO, J. F., "Artículo 105" in RODRÍGUEZ-PIÑERO Y BRAVO-FERRER, M. & CASAS BAAMONDE, M. E., (dirs.), *Comentarios a la Constitución Española*, *Tomo II*, Madrid, Fundación Wolters Kluwer, Boletín Oficial del Estado, Tribunal Constitucional y Ministerio de Justicia, 2018, pp. 487-501.
[1515] Act 19/2013, 9th November, on transparency, access to public information and good government.

which are in the possession of any [public administration or assimilated entity] and which have been drawn up or acquired in the exercise of their functions". Article 14 establishes a series of limits to the right of access to public information which, similarly to the limits set by the GDPR regarding transparency-related rights, include intellectual property, national and public security.

The relevant public body (the Ministry for Ecological Transition) did not respond to the access request. The implicit refusal to disclose the requested content was appealed before the Spanish Transparency Council. Within said procedure, the Ministry for Ecological Transition explicitly refused to disclose the algorithm's source code alleging national and public security risks and the conflict with intellectual property rights. The Transparency Council rejected the national and public security risk argument, which were not even developed.[1516] The Council did however partly accept the argument that the source code was protected by intellectual property and therefore denied CIVIO access to the full source code. It did however, command the Ministry to disclose the programme's technical specifications; the result of the tests carried out to check that the system complied with the functional specification; and, any other deliverable that conveys knowledge on how the application operates.[1517]

In other words, the information released comprehended the objectives and general commands that the system had to consider and respect when making decisions, as well as the tests that were carried out in order to confirm that the programme worked in the way it was supposed to. However, the actual document in which the exact and specific instructions that the system followed when it was executed were not revealed as it was considered that this particular information fell under the scope and was protected by intellectual property rights. This resolution has been appealed before the Spanish Central Administrative Court, the decision is still pending.

Perhaps the most striking element in this process is the fact that an algorithm that was fully developed by a public administration and was used to deliver administrative decisions, was considered to be protected by intellectual property rights. In this sense, it is relevant to

[1516] Spanish Transparency and Good Government Council, Resolution 701/2018, February 18th, 4th legal foundation.
[1517] Spanish Transparency and Good Government Council, Resolution 701/2018, February 18th, resolution.

highlight that article 13 of the Spanish Intellectual Property Act[1518] excludes the following documents from its scope of protection: laws; regulations; draft legislation of all types; court decisions; public body resolutions, agreements, deliberations, opinions; and, all official translations of the aforementioned texts. Drawing from this provision, it is relevant to recall that the programme made the decisions on whether an applicant met the requirements to benefit from this particular form of public aid. Thus the algorithm itself constituted a public body resolution, agreement, deliberation and opinion. Additionally, since the criteria the algorithm applied differed from the requirements set out by the regulation on the social bond, the source code of the algorithm, that is the text containing the specific instructions that the algorithm applied, effectively constituted the actual text of the social bond regulation. Hence, intellectual property rights do not protect this information.

Moreover, as the appellants in this case have argued, it does not make sense for public administrations to constantly disclose all types of information that could actually fall under the protection of intellectual property rights, as it is the case in Spain, but apply those limits to the use of algorithms.[1519] The application of intellectual property rights to public production must only take place in very specific cases, for the public interests that these rights conflict with, are essential in the construction of democratic societies.

It seems fairly obvious that the issue here lies in the inability that regulators and public institutions seem to have in coming to terms with the clear analogies that exist between traditional forms of regulatory production and administrative decision-making and the new forms that are mediated by or rely exclusively on automated systems and that should therefore be subjected to the same transparency and other general requirements as traditional forms of decision-making.[1520] In this case, the algorithm contained the general rules that were (erroneously) applied to specific cases and, consequently, effectively acted as a regulatory instrument. The algorithm should therefore be fully transparent in order to comply with the publicity requirements applied to regulatory instruments in democratic countries. The regulatory quality of algorithmic systems will be further argued in the following chapter.

Moreover, requiring that public algorithms are fully transparent enhances public accountability, as it is possible to determine the exact (and true) logic behind a decision.

[1518] Royal Legislative Decree 1/1996, 12th April 1996, which passes the revised text of the Law on Intellectual Property, regularising, clarifying and harmonising the legal provisions in force on the subject.
[1519] DE LA CUEVA, J., "El derecho a no ser gobernados mediante algoritmos secretos", *cit.*, 2019.
[1520] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, pp. 223-269.

However, for said enhanced accountability to be realised it is necessary to develop specific regulatory instruments applicable to the public sector use of algorithms. Otherwise, there is a very real risk of normalising the control of public sector automated decision-making through private instruments that, in turn, leads to a decrease in the requirements public bodies must meet when developing their activities and the guarantees that citizens are provided, especially when public institutions exercise powers that can affect the fundamental rights of individuals.[1521]

3.1.2. Administrative courts granting transparency

The Administrative court of Lazio-Roma, Section III bis, established in Judgment No. 3769 of 22nd March 2017 that the plaintiff should have full access to the algorithm used in order to manage teacher mobility as it determined the existence of a "direct, concrete and current interest corresponding to a legally protected situation and linked to the document to which access [was] requested". It also established that this transparency mandate was not limited to general explanations but to the full source code of the software used.

The obligation to disclose algorithms used by public administration bodies was also established by the Catalan Commission for the Guarantee of the right of access to public information, as it considered that confidentiality and the secrecy of decision-making processes could not serve as a legitimate basis to prevent the disclosure of an algorithm used in selecting the teachers tasked with marking university access exams.[1522]

What is particularly worrying in these cases is the failure of public bodies to comply with the most basic transparency requirements in non-conflictive decisions, which brings up questions of the extent to which public administrations are actually aware of the risks that these systems can pose in some cases but also of their mandate to serve public interests.

3.1.3. Banning the use of algorithms in the public sector: the Dutch "SyRI" case

The recent ruling[1523] on the "SyRI" system in the Netherlands has become widely discussed as it touches upon a series of very relevant elements regarding the use of algorithms in general and, more specifically, by public administrations. The "SyRI" system was used by

[1521] *Idem*, pp. 248-249.
[1522] Catalan Commission for the Guarantee of the right of access to public information, Joined decisions 123/2016 and 124/2016.
[1523] District Court of The Hague, ruling of February 5th 2020, case number C / 09/550982 / HA ZA 18-388.

Dutch public bodies in order to detect and prevent welfare fraud. This system, which was not particularly effective, helped to perpetuate the dominant narratives that stigmatise the poor and welfare recipients.[1524]

In this case, the regulation that served as legal basis for the automated system was claimed by the plaintiffs to be in breach of article 8 of the European Convention on Human Rights as well as other provisions contained in the Charter of Fundamental Rights of the European Union, the International Convention on Civil and Political Rights and the GDPR.

The Court considered that the aim for which the system was being used, that is, welfare fraud detection, was legitimate. However, the way in which massive amounts of personal structured data were entered into the algorithm, the system's self-learning capacity and the lack of information rights and general transparency of the system were not sufficiently supported by the regulatory instrument that authorised its implementation, as it did not sufficiently protect against possible arbitrary outcomes nor adequately determined the degree of discretion granted to authorities with regard to the way in which the system was used and its effects evaluated.[1525] Additionally, the Court considered that there was no "fair balance" between the goals of the system and its interference in the private life of individuals, particularly considering the safeguards placed were insufficient and therefore failed to provide an adequate justification for said interference.[1526]

In this vein, it is important to take into consideration that the actions of public bodies and, thus, of the automated systems employed by said bodies, are constrained by the limits imposed by the principle of legality, especially in civil law systems. Unlike private parties, the actions of public administrations are not subjected to the principle of autonomy but to the principle of legality. This principle requires public bodies to act within the limits set by the laws and regulations that are applicable to them. Moreover, especially when it comes to activities that entail interfering in the fundamental rights of individuals, the material perspective of this principle reserves the basic regulation of these issues makes it necessary for an Act passed by parliament to contain a series of sufficiently detailed rules that serve as legal basis for the actions of public bodies. The legal instruments that support public interference in the fundamental rights of individuals must clearly establish the scope and

[1524] RANCHORDAS, S. & SCHUURMANS, Y., "Outsourcing the welfare state…", *cit.*, 2020, p. 6.
[1525] District Court of The Hague, ruling of February 5th 2020, case number C / 09/550982 / HA ZA 18-388, paragraphs 6.66-6.70.
[1526] *Idem*, paragraphs 6.80-6.107.

conditions of the interference as well as sufficient safeguards to guarantee the protection of individuals' rights. In this case, the Court concluded that the authorisation provided by law was too wide and open to sufficiently limit the actions of public institutions in the deployment of "SyRI".

It is also particularly significant that, while the Court did consider several GDPR provisions in its ruling, including the personal data processing principles of article 5 and was highly critical with the state's failure to conduct the mandated DPIA of article 35, it mainly based its ruling on article 8.2 of the European Convention on Human Rights, which specifically establishes the requirements that public authorities must comply with in order to interfere with the private and family life of individuals.

This particular choice raises questions on the effectiveness of the GDPR in regulating the use of automated systems by public bodies. As it has already been stated, the GDPR is mainly structured as a regulatory instrument aimed to be applied to private parties. Moreover, the provisions concerning data processing by public authorities seem to be simply thrown into the mix of hard and soft law without specific consideration to the particular nature of public institutions. This unavoidably limits the applicability of the Regulation in cases such as the one here examined in which a classical but comprehensive proportionality test had to be carried out.

4. THE SHORTCOMINGS OF ACCOUNTABILITY MECHANISMS

The insufficiencies of the individual rights approach render the existence of an overarching system of algorithmic governance necessary. As the previous chapter conveyed, the GDPR contains a series of instruments that are aimed towards governing algorithms and achieving system transparency and accountability. This system however fails to effectively control the design and development of algorithms and protect individuals against the harms they may cause.

The tools for governing algorithmic systems include rule-setting mechanisms, on-going control mechanisms and enforceability mechanisms, also known as mechanisms to ensure enforceability. The GDPR includes all of the above stated mechanisms but there are also other rules and proposals put forward both by public institutions and the scholarship that aim

to set up a comprehensive governance system that rules over and controls the use of algorithms.

The GDPR's shortcomings in this respect do not just result from its focus on privacy but mainly from the Regulation's excessive reliance on the actions of controllers to implement the rule-setting and on-going control mechanisms and its subsequent lack of enforceability.[1527] In the same way that the individual rights regime in privacy instruments is structured around the notion of consent and therefore excessively relies on and trusts individuals' ability to understand and exercise the available rights for data protection, the GDPR's system of what KAMINSKI labels "collaborative governance",[1528] relies too heavily on the actions of controllers to implement the oversight tools to ensure compliance with the Regulation.[1529]

Firstly, the implementation of codes of conduct and certification mechanisms is voluntary.[1530] Moreover, even when private parties decide to develop these mechanisms, there are still significant shortcomings to the level of accountability that they can achieve. Neither codes of conduct or certification mechanisms require third-party participation in their development. In the case of codes of conduct, their development is set up as a dialogue between authorities and "associations or other bodies representing categories of controllers and processors,"[1531] which means that said instruments will be skewed, favouring the interests of controllers and processors and not consider third party interests that will result affected by data processing activities.

Secondly, the technical standards specified in the "data protection by design and by default"[1532] mandate require controllers and processors to carry out their data processing activities with regard to the principles set in article 5 of the GDPR. These requirements do not pose significant limits on the actions of controllers and processors seeing as they are quite general and vague and will therefore probably allow for many types of processing activities

[1527] KOOPS, B. J., "The problem with European data protection law", *cit.*, 2014, pp. 253-256; SCHREURS, W. *et al.*, "Cogitas, ergo sum…", *cit.*, 2008, p. 268.
[1528] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1529-1616.
[1529] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1607-1610; KOOPS, B. J., "The problem with European data protection law", *cit.*, 2014, pp. 253-256.
[1530] Articles 40 and 42 GDPR.
[1531] Article 40.2 GPDR.
[1532] Article 25 GDPR.

to be considered necessary and justified and therefore falling within the scope permitted by articles 25 and 5 of the GDPR.[1533]

Compliance with the data protection mandate could be argued to be partly accomplished in those cases in which data protection impact assessments (DPIAs) are mandatory[1534] seeing as the obligation to carry out an impact assessment will force controllers to consider what data they are collecting and processing.[1535] However, it is inevitable to show certain scepticism towards the effectiveness of DPIAs. DPIAs are not always mandatory and the cases in which they are, are determined through the undetermined legal concept of "high risk" which, depending on how the notion is implemented, may provide controllers and processors with a wide range of cases in which to circumvent this obligation.

More importantly, the current governance system set up by the GDPR does not seem to be capable of ensuring that controllers and processors carry out impact assessments. The Dutch SyRi system, which was earlier mentioned, is a clear case in point, seeing as the ruling indicated that, while a general DPIA was carried out, such assessment was carried out before the entry into force of the GDPR and individual DPIAs were not implemented for each different project for which the SyRI system was deployed, thereby failing to meet the requirements set by article 35 of the GDPR.[1536] Therefore, if public authorities are not meeting the mandatory requirements set by the governance toolkit in the GDPR, the likelihood that private parties will comply and effectively apply said accountability and control mechanisms must certainly be put into question. In this sense, since the GDPR does not require DPIAs to be made public, it heavily curtails the chances of third-party oversight, which is the very essence of impact assessments.[1537]

Moreover, since the regulation of DPIAs contained in the GDPR is quite general, controllers may be able to comply with impact assessment requirements by simply establishing checklists that are not really effective in detecting compliance with "data protection by design

[1533] KOOPS, B. J., "The problem with European data protection law", *cit.*, 2014, pp. 254-255.
[1534] Article 35.1 GDPR: "Where a type of processing in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data".
[1535] KOOPS, B. J., "The problem with European data protection law", *cit.*, 2014, p. 255.
[1536] Court of The Hague, ruling of February 5th 2020, case number C / 09/550982 / HA ZA 18-388, paragraphs 6.104-6.105.
[1537] KAMINSKI, M. E., "Binary governance…", 2019b, *cit.*, pp. 1608-1609.

and by default" technical standards.[1538] This is a clear example of the difficulties of combining traditional forms of regulation with algorithmic design. Since the GDPR is applicable to all private economic sectors and public institutions (with the exception of law enforcement), the regulation of DPIAs must be general in order to allow for its specific implementation in all the different areas it covers.

However, by entrusting controllers with the power that the specific implementation of technical standards required by the GDPR and accountability mechanisms such as DPIAs entail, the Regulation provides said actors with an ample range of action in which to decide the extent to which and how they want to comply with said mandates. Hence, unless more specific requirements are set up over time by public bodies, it will be up to controllers to decide whether they want to fully comply with personal data protection mandates or to develop their actions in a way in which there is formal compliance but no real integration and inclusion of data protection mandates in the organisations activities.[1539]

Finally, another difficulty of implementing control mechanisms by controllers lies in the complexity of the GDPR. As it has already been discussed, the GDPR has a very wide scope of application and aims to cover all types of processing activities carried out by both the private and public sector with the exception of the types of activities that fall within the scope of application of the law enforcement Directive. The lack of sector-specificity and the length and complexity of the rules means that implementing the Regulation's provisions is a difficult and costly job. Consequently, it is only logical that controllers and processors will prefer to focus on mandatory provisions rather than on the voluntary accountability mechanisms included in the GDPR.[1540]

All in all, the European legislator has largely failed in establishing an effective algorithmic governance system for personal data protection (not to speak of the protection of other human rights such as the right to equality and non-discrimination). This lack of effectiveness can

[1538] KOOPS, B. J., "The problem with European data protection law", *cit.*, 2014, p. 255: "I fear that, as long as data protection is not in the hearts and minds of data controllers—and the law so far has done a poor job in reaching those hearts and minds … —mandatory data protection impact assessments will function as paper checklists that controllers duly fill in, tick off, and file away to duly show to auditors or supervisory authorities if they ever ask for it. Procedure followed, problem solved. But truly following an impact assessment, and particularly translating its findings into actual data protection- friendly designs and default settings … requires an attitude that looks beyond legal compliance, a vision that sees the logic behind the many legal rules of data protection, a mindset that is aware of the rationale of data protection".
[1539] KOOPS, B. J., "The problem with European data protection law", *cit.*, 2014, p. 255
[1540] *Idem*, pp. 254, 256-258.

largely be linked back to the fact that the GDPR does not set up a mandatory governance framework, seeing as most of the governance instruments are voluntary. In addition, even when these tools are mandatory or when controllers choose to implement them, the degree of enforceability is very low.

One of the main reasons for this failure in enforceability results from the elimination of third-party stakeholders from the governance equation. Governance partnerships between the actors carrying out risky activities and the authorities in charge of supervising and applying the relevant regulatory instruments necessarily need the intervention of organisations that represent interests affected by said activities. If the accountability tools set up are to be deployed by the actors carrying out risky activities (controllers), the participation of third parties can only be granted through real and effective transparency, which is currently largely missing. In conclusion, the GDPR aims to set up a collaborative governance framework without one of its main pillars.[1541]

5. THE RELATIONSHIP BETWEEN PERSONAL DATA PROTECTION, EQUALITY AND NON-DISCRIMINATION

5.1. THE PRIVACY VS. ANTIDISCRIMINATION DILEMMA

Neither the GDPR nor Directive 2016/680 for data protection in law enforcement are built as antidiscrimination instruments. The objective of the European regulatory framework is to protect and enact the fundamental right to data protection, not the right to equality and non-discrimination.

This does not mean that preventing and fighting discrimination is not amongst the objectives of said regulatory instruments. For example, Recital 71 of the GDPR does state that controllers should "implement technical and organisational measures appropriate" to prevent discriminatory effects, Recital 23 of Directive 2016/680 for data protection in law enforcement establishes a general prohibition of discrimination based on genetic features and Recital 38 of the Directive sets out a general prohibition regarding profiling "that results in discrimination against natural persons on the basis of personal data which are by their nature particularly sensitive in relation to fundamental rights and freedoms".

[1541] KAMINSKI, M. E., "Binary governance…", *cit.*, 2019b, pp. 1608-1610.

Additionally, Recital 75 of the GDPR and 51 of the Directive also convey the awareness and general concern regarding risks of discrimination resulting from the processing of personal data and Recital 85 of the GDPR and 61 of the Directive set a series of obligations for controllers in order to deal with the discriminatory risks that might result from data breaches.

The general prohibition regarding the processing of special categories of personal data and the specific and reinforced prohibition on making decisions based solely on automated processing of special categories of data also show the extent to which the European regulatory framework aims to protect individuals against algorithmic discrimination. This notion is reinforced in the Directive by specifically prohibiting "profiling that results in discrimination against natural persons on the basis of special categories of personal data".

Since the main objective of existing regulatory instruments is to protect informational privacy rights, anti-classification serves as the ideal policy instrument to provide some protection against discrimination while mainly focusing on privacy. The case for privacy or anti-classification rests on the notion that members of disadvantaged groups have generally been discriminated against when decision-makers became aware of their protected group membership.[1542] Since, as it was already explained in part I, algorithms are easily and heavily influenced by existing biases in society, they will incorporate hierarchical structures that subordinate certain groups and result in discriminatory decisions regarding members of said groups. This is the general argument wielded by privacy advocates who consider that by hiding certain information that will reveal individuals' membership to disadvantaged groups, their risk of being discriminated will be reduced.[1543]

5.1.1. Less information can lead to wrong inferences

STRAHILEVITZ[1544] offers an opposing view to the one adopted by the proponents of anti-classification as he argues that reducing privacy and offering decision-makers more access to an individual's personal information will reduce the chances of decisions being made largely based on profiling. The example he offers is companies not wanting to hire ex-convicts. In this case, race operates as a proxy for being an ex-convict which employers use whenever

[1542] ROBERTS, J. L., "Protecting privacy to prevent discrimination", *cit.*, 2015, pp. 2099.
[1543] ŽLIOBAITĖ, I. & CUSTERS, B., "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models", *cit.*, 2016, p. 188.
[1544] STRAHILEVITZ, L., "Privacy versus antidiscrimination", *cit.*, 2008, pp. 363-381.

they do not have access to the applicants' criminal records.[1545] In fact, HOLZER, RAPHAEL and STOLL,[1546] whose paper STRAHILEVITZ cites, proved that employers who carried out criminal background checks were much more likely to hire black Americans, especially male, than those who did not.[1547] The theory proposed by STRAHILEVITZ was recently confirmed by DOLEAC and HANSEN,[1548] who conclude that black and Hispanic men have lower chances of finding a job when there is a ban on employers accessing criminal histories of potential employees, seeing as they use race as a proxy for past convictions.

Although the cited articles refer to decisions made by humans, the idea that a trade-off between privacy and discrimination could take place can also be applicable to automated processing used in cases in which decisions that affect individuals are made. One of the risks of algorithms is that they will generate profiles and predictions that will reflect existing social discrimination and that will apparently confirm prejudices held against certain groups.[1549] This is precisely, what happens in the particular case above cited. African American men are largely overrepresented in the American prison system due to structural discrimination in law enforcement and the criminal justice system.[1550] Consequently, when the employer decides to hire less Black men she is basing her decision not on false notions such as "black people are lazy", but on the actual statistical truth that within the pool of Black men the probability that some of them are ex-convicts is higher than within other races or within the pool of female candidates.[1551]

As the following section will further explain, even if the processing tool is not directly fed the candidate's race, it is very likely that it will infer it from other data,[1552] which will probably lead to the algorithm predicting African American male applicants to be ex-convicts at a higher rate than individuals belonging to other ethnic groups or women. Consequently, feeding the algorithm the actual information on criminal records would benefit African American applicants with no past convictions. However, this would undoubtedly result in

[1545] *Idem*, p. 366.
[1546] HOLZER, H. J., RAPHAEL, S., & STOLL, M. A., "Perceived criminality, criminal background checks, and the racial hiring practices of employers", *cit.*, 2006, pp. 451-480.
[1547] *Idem*, p. 452.
[1548] DOLEAC, J. L. & HANSEN, B., "The unintended consequences of 'ban the box'…", *cit.*, 2020, pp. 321-374.
[1549] See, for example, the case of the recidivism risk algorithm created by Northpointe described in Chapter IV.
[1550] BERTRAND, M., MULLAINATHAN, S. & ABRAMS, D., "Discrimination in the judicial system", *cit.*, 2001; GROSS, S. R., POSSLEY, M. & STEPHENS, K., "Race and wrongful convictions in the United States", *cit.*, 2017.
[1551] STRAHILEVITZ, L., "Privacy versus antidiscrimination", *cit.*, 2008, p. 366.
[1552] ŽLIOBAITĖ, I. & CUSTERS, B., "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models", *cit.*, 2016, p. 185.

worse consequences for those particular job applicants who had served time in prison. Policymakers and regulators must therefore, in this particular case, decide whether it is more valuable to protect racial minorities or reinsert individuals with criminal records in society. It is important to highlight that not all cases will create such complicated trade-offs but that, what this comes to prove, is that a general ban on accessing certain pieces of data may in some cases be counterproductive.

5.1.2. Anti-classification does not prevent indirect algorithmic discrimination

The processing power that machine learning algorithms have, allows for correlations to be established even between pieces of data that are apparently unrelated.[1553] Hence, protected group membership can be inferred from other pieces of data. In addition, as machine learning algorithms become more powerful and are able to withdraw correlations from larger datasets it will be very difficult to figure out how certain correlations have been drawn or to even detect them.[1554]

Consequently, as it was already mentioned with regard to article 9 of the GDPR it is very difficult to effectively enforce provisions that prohibit the processing of personal data that could reveal protected group status, since virtually all data is, when combined in some form, susceptible of producing such predictions. Even if approaches such as the "sensitive-by-distance" [1555] proposal, which aims to set rules to determine what data could realistically, when analysed, provide information on special categories of personal data, are adopted, it is still unlikely that all information that could potentially reveal special categories of personal data will be detected and included within the prohibition. Especially because in many cases data controllers and processors will not even be aware of the fact that some of the data fed into the algorithm could reveal sensitive information.

[1553] YOUNG, E., "Educational privacy in the online classroom…", *cit.*, 2015, p. 550.
[1554] GOODMAN, B. & FLAXMAN, S., "European Union Regulations on algorithmic decision making and a "right to explanation", *cit.*, 2017, pp. 53-54: "Furthermore, as datasets become increasingly large, correlations can become increasingly complex and difficult to detect. The link between geography and income may be obvious, but less obvious correlations—say between IP address and race—are likely to exist within large enough datasets and could lead to discriminatory effects. For example, at an annual conference of actuaries, consultants from Deloitte explained that they can now 'use thousands of 'non-traditional' third party data sources, such as consumer buying history, to predict a life insurance applicant's health status with an accuracy comparable to a medical exam'"
[1555] MALGIERI, G. & COMANDÉ, G., "Sensitive-by-distance…", *cit.*, 2017a, pp. 238.

As many studies have shown, omitting protected characteristics as input variables in algorithms does not help to prevent discrimination.[1556] For example, GILLIS and SPIESS find that, even if the number of variables that correlate with race are reduced and excluded from the dataset fed into the algorithm, there is still significant disparity amongst different ethnic groups when predicting risk of loan default.[1557] While this anticlassification mechanism will undoubtedly hamper the chances of direct discrimination taking place, as no direct consideration of protected variables will be included, the large amounts of available data lead to the practical impossibility of preventing indirect discrimination through this mechanism. Moreover, in some cases, direct discrimination by inference will take place and could be mistaken for instances of indirect discrimination which allow for justifications to be provided. Algorithms can make decisions based on variables or combinations of variables which act as proxies for special categories of data, hence leading to instances of discrimination for individuals in which said categories concur.[1558]

5.1.3. Anti-classification through privacy does not solve group disadvantage and can reinforce it

Even when hiding protected characteristics might be a solution to prevent some cases of discrimination, the privacy approach does not solve the larger structural problem due to which some groups in society are subordinated to others.[1559] In fact, making individuals hide such important aspects of their personal identity could lead to members of disadvantaged groups internalising that there is something actually negative or shameful about those characteristics.[1560]

One of the main problems that results from structures of hierarchy within society is that the particularities of groups that have undergone historical oppression are not taken into account and that most power structures and services are designed based upon the prototypical notion of the liberal individual. Therefore, if the processing of data that takes into account categories of data that identify disadvantaged group membership is completely forbidden, it could be

[1556] WILLIAMS, B. A., BROOKS, C. F. & SHMARGAD, Y., "How algorithms discriminate based on data they lack…", *cit.*, 2018, pp. 78-115.
[1557] GILLIS, T. B. & SPIESS, J. L., "Big data and discrimination", *The University of Chicago Law Review*, vol. 86, No. 2, 2019, pp. 469-470.
[1558] ŽLIOBAITĖ, I. & CUSTERS, B., "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models", *cit.*, 2016, p. 185.
[1559] ROBERTS, J. L., "Protecting privacy to prevent discrimination", *cit.*, 2015, p. 2124.
[1560] *Idem*, p. 2160.

very easy to fall back on using said prototypical individual instead of drawing actually useful insights regarding the particular characteristics and needs of members of disadvantaged groups in order to reframe society towards the creation of more inclusive structures.

By not paying attention to existing structures of subordination and disadvantage and failing to introduce the particularities of protected groups into automated processing tools, the disadvantage suffered by members of these groups will probably be reinforced given the fact that the inputs used in all types of decision-making are mainly shaped according to the characteristics of dominant groups. For example, GILLIS and SPIESS argue that, since minorities are more likely to apply for loans to finance entities that are not mainstream lenders, if credit scoring companies evaluate said loans less favourably, biases against individuals that belong to minority groups will be introduced in credit scores that will then determine or in some way influence future credit applications made by said individuals.[1561] By introducing special category membership as a variable and providing the algorithm with the relevant instructions, it will be able to correct for said biases.[1562] With regard to this idea, if affirmative action mechanisms were to be undertaken in algorithmic decision-making in order to compensate for existing structural biases in society, including sensitive data regarding disadvantaged group membership would be necessary.[1563]

Moreover, if protected group membership is not included, the problem with having access to such a huge volume of data is that it is easier for decision-makers to decide based on apparently harmless characteristics which may in fact hide a discriminatory rationale.[1564] In this sense, part of the literature has argued in cases in which discrimination actually occurred it would be easier to detect it if protected class membership was included in the dataset. [1565]

Thus, in general, detecting and fixing discriminatory outcomes will be much easier if protected characteristics are included in the decision-making process.[1566] In order to ensure that the benefits of considering disadvantaged group membership in decision-making can be reaped, CUSTERS and ŽLIOBAITĖ propose including an exemption in the European legal

[1561] GILLIS, T. B. & SPIESS, J. L., "Big data and discrimination", *cit.*, 2019, pp. 471-472.
[1562] *Ibidem.*
[1563] ROBERTS, J. L., "Protecting privacy to prevent discrimination", *cit.*, 2015, p. 2157.
[1564] BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *cit.*, 2016, p. 692; STRAHILEVITZ, L., "Privacy versus antidiscrimination", *cit.*, 2008, pp. 373-374.
[1565] GOODMAN, B., "A step towards accountable algorithms?...", *cit.*, 2016,p. 3.
[1566] FELDMAN, M. *et al.*, "Certifying and removing disparate impact", *cit.*, 2015; GOODMAN, B., "A step towards accountable algorithms?...", *cit.*, 2016, p. 3.

framework that allows the processing of special categories of personal data when said processing is carried out by models that are designed and aimed towards reducing discrimination.[1567]

5.2. COMBINING THE ANTI-DISCRIMINATION AND DATA PROTECTION FRAMEWORKS

The shortcomings that the privacy and anti-discrimination frameworks individually considered present and the search for a way in which to address algorithmic discrimination with existing regulatory instruments have led several scholars to suggest combining both systems in order to deal with instances of discrimination generated by automated systems.[1568]

One of the main problems with the anti-discrimination framework is the fact that the enforcement of prohibitions to discriminate can only be carried out after the discriminatory action has taken place. In addition, access to justice can be limited due to a number of factors when it comes to defending the rights to equality and non-discrimination. On the one hand, some cases of discrimination are very hard to detect and even harder to prove by their victims. This is an issue that, as a result of system opacity, is particularly present when dealing with instances of algorithmic discrimination, especially considering the opportunities of justifying instances of discrimination.[1569] On the other hand, many victims of discriminatory situations are in vulnerable positions that hinder their access to the means, knowledge and mechanisms that are necessary in order to bring forward cases of discrimination.[1570] All in all, the individualistic basis upon which the structure of judicial remedies for cases of discrimination is built upon fails to grasp the full nature of discrimination and places an excessive burden on individuals who are generally in already socially disadvantaged and vulnerable positions.

The European data protection framework offers a series of instruments that, to a certain extent, fill the gaps of the anti-discrimination framework when it comes to dealing with algorithms and the risks to the right to equality that they generate.[1571] The GDPR and other

[1567] ŽLIOBAITĖ, I. & CUSTERS, B., "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models", *cit.*, 2016, p. 198.
[1568] DRECHSLER, L. & BENITO SÁNCHEZ, J. C., "The price is (not) right…", *cit.*, 2018, pp. 1-23; HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, pp. 1143-1186.
[1569] XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *cit.*, 2020, p. 174; HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, pp. 1167-1168.
[1570] XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence…", *cit.*, 2020, pp. 174-175.
[1571] HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, pp. 1170-1171.

data protection legislation offer the possibility of carrying out an *ex ante* control of automated systems and reinforce the *ex post* control offered by anti-discrimination legal instruments. The principles contained in article 5 of the GDPR force controllers and processors to ensure that the input data is not biased (accurate) and that data is processed is "lawfully, fairly and in a transparent manner", which indirectly introduces the obligation to at least consider how the way in which a system processes data affects the rights to equality and non-discrimination.[1572]

More importantly, the transparency principle is backed up by the due process rights and accountability mechanisms contained in the Regulation. The system of individual rights recognised in the GDPR could provide the elements necessary for individuals to gain awareness of the discriminatory automated decisions they might be subjected to and prove some cases of algorithmic discrimination. In addition, if properly implemented, the accountability and governance mechanisms such as codes of conduct, certification and DPIAs could not just help to bring about an *ex ante* and on-going control of algorithmic systems but also help fill in the shortcomings of systems for the protection of fundamental rights that mainly place the burden on the individual.

Article 24.1 of the GDPR also offers the backdrop against which to develop the "equality by design mandate". Said provision indicates that:

> "Taking into account the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons, the controller shall implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation. Those measures shall be reviewed and updated where necessary."

Article 24 must be analysed in relation to article 1.2 of the GDPR, which establishes that the GDPR protects the fundamental rights and freedoms of natural persons, amongst which are the rights to equality and non-discrimination. Additionally, recital 71 establishes that controllers should introduce measures in order to prevent discriminatory effects based on special categories of data and recitals 75 and 85 establish the possibility of discrimination as an undesirable risk of data processing.

[1572] *Idem*, p. 1172.

Moreover, the antidiscriminatory goals of the GDPR are also conveyed through the prohibition to process special categories of data contained in article 9 of the Regulation and the stricter requirements set by article 22, which heavily limits the possibility of basing solely automated decision-making, including profiling, on the special categories of data included in article 9.[1573]

Consequently, article 24.1, above cited, provides the necessary framework that justifies the need for controllers to introduce certain technical elements into algorithms that ensure that they do not produce discriminatory results. We therefore consider that, just like "data protection by design and by default"[1574] is mandated by the GDPR, so is the notion of "equality by design".[1575]

One of the shortcomings of the equality and non-discrimination framework that was earlier pointed out was the fact that it is not clear whether search engine operators can be accused of discriminatory practices when reproducing negative stereotypes of disadvantaged groups. However, an alternative pathway may be opened through article 9 of the GDPR as the CJEU has established that the prohibition of processing special categories of personal data does apply to search engine operators.[1576]

In this sense, whenever algorithmic discrimination is the result of data collection or processing activities that do not fully comply with the data protection framework, victims of discrimination may also base their claims on the fact that the data was illegally obtained or processed.

Even with its many shortcomings and the tensions between privacy and anti-discrimination frameworks, it is undeniable that the GDPR can fill in some of the voids left by anti-discriminatory legal instruments and vice versa and that together they do offer a more comprehensive system of protection against algorithmic discrimination. The combination of both legal frameworks offers the blueprint upon which to base the very needed improvements that the current regulatory system needs in order to properly address algorithmic discrimination. However, for the time being, and given the obvious lagging in innovative

[1573] DRECHSLER, L. & BENITO SÁNCHEZ, J. C., "The price is (not) right…", *cit.*, 2018, pp. 8-9.
[1574] Article 25 of the GDPR.
[1575] HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, pp. 1172, 1179.
[1576] CJEU Judgment 24th September 2019, C-136/17, GC and Others v. Commission nationale de l'informatique et des libertés (CNIL).

institutional and legal arrangements to control for all the risks posed by algorithms and, in particular, algorithmic discrimination, technical and legal operators should focus on combining the data protection and equality and anti-discrimination frameworks in order to achieve a more effective system that protects the rights to equality and non-discrimination against the hazards generated by automated decision-making.

CHAPTER IV. POSSIBILITIES AND PROPOSALS FOR THE REGULATION OF ALGORITHMS

This chapter, as well as most of the dissertation, is built from an idea that has been clearly expressed by SHNEIDERMAN as "in software, things often go bad".[1577] It is true that algorithms may sometimes be more accurate than humans and that they can be tested and experimented with, but it is also true that automated systems are created by humans that can inadvertently or purposefully introduce biases into the algorithm, that these systems can make mistakes and that the degree of complexity and lack of intuitiveness of the results yielded by algorithms can cause significant harms, in particular, to the rights of vulnerable members of society.

Up until now the dissertation has focused on examining the ways in which existing regulatory instruments can deal with algorithmic discrimination and other issues and problems that arise from the use of data processing technologies. This chapter examines some of the proposals that have been put forward with regard to the development of new regulatory instruments and bodies aimed to control algorithms. The objective is to combine some of the regulatory proposals in order to put forward a comprehensive system to regulate algorithms with a special focus on the risks automated systems generate for the equal treatment principle in all its expressions.

This chapter is divided in to four sections. The first and second sections focus on considerations regarding the regulation of algorithms in the public and private sectors respectively. The third section draws a series of consideration on the widespread debate on algorithmic transparency. The final section puts together a series of proposals for a system of public regulation and intervention to prevent and deal with the risks and harms caused by algorithms.

1. CONSIDERATIONS REGARDING THE REGULATION OF ALGORITHMS EMPLOYED BY PUBLIC ADMINISTRATIONS

In order to establish a framework that regulates the use of public automated decision-making it is necessary to analyse and determine how the object of regulation should be classified

[1577] SHNEIDERMAN, B., "Algorithmic Accountability: Designing for safety through human-centered independent oversight", *Turing Lecture (The Alan Turing Institute)*, 31st May 2017. Available on 23rd May 2020 at: https://www.youtube.com/

from a legal perspective. This section aims to answer the following question: what is the legal nature of algorithms used by the public sector?

The debate surrounding the public use of automated decision-making systems focuses on two key issues. Firstly, on whether algorithms have legal effects and, secondly, if they are determined to have legal effects, whether they should be classified as regulatory instruments, individual administrative decisions or as something altogether different. If algorithms are deemed to have legal effects, the requirements they should be subjected to are higher and stricter and citizens should be offered greater guarantees and protections than if automated systems are considered simple software programmes used to assist in administrative procedures.

Following the work of LESSIG,[1578] YEUNG[1579] and BOIX PALOP,[1580] this section approaches the regulatory nature that algorithmic systems have when used by public bodies as elements that influence, shape and predetermine human behaviour and which sometimes even establish binding decisions, obligations or prohibitions.

1.1. ALGORITHMS USED BY PUBLIC ADMINISTRATIONS ARE LEGAL INSTRUMENTS

Some scholars and public administrations have defended that algorithms used in administrative procedures constitute no more than software programmes either designed to execute previously made decisions or to serve as support tools in administrative procedures. In both cases, they argue, these systems lack any legal effect. However, an increasing number of public decisions are being delegated to automated systems that establish the final resolution that affects the individual recipients of administrative decisions.

In this sense, the Administrative court of Lazio-Roma[1581] ruled that an algorithm used by the Italian Ministry of Education to manage teacher mobility had legal effects. The Italian Ministry of Education defended that automated systems are mere software programmes that have no legal implications whatsoever, whereas the plaintiff considered the algorithm to have legal effects and hence, to effectively be an administrative decision. The objective of the plaintiff's claim was to gain full access to the algorithm.

[1578] LESSIG, L., *Code: Version 2.0*, New York, Basic books, 2006.
[1579] YEUNG, K., "Algorithmic regulation…", *cit.*, 2018, pp. 505-523.
[1580] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020.
[1581] Administrative Regional Court of Lazio-Roma, Section III bis, Judgment No. 3769, 22nd March 2017 (own translation).

The Ministry of Education argued that the actions of these systems are limited to executing administrative decisions (resolutions) that have already been made. The Ministry used this argument in order to support its claim that algorithms should not be subjected to the transparency requirements applicable to administrative documents.

However the court considered that the algorithm did have legal effects and, in fact, constituted an administrative decision as it stated that:

> "The algorithm ends up, therefore and in the end, providing the basis for the procedure mentioned above, since the identification, in concrete terms, of the specific location of each individual teacher in the context of mobility is identified exclusively by the algorithm mentioned above;
>
> The endoprocedural acts of collection of the data necessary for the purposes of the procedure, as well as the final act of the procedure itself, have been merged and exhausted only in the operation of the algorithm in question, with the further consequence that the assimilation of the algorithm in question to the administrative act can and must be considered…"[1582]

Algorithms are being used in public decision-making procedures that affect individuals and, even when their objectives are limited to implementing pre-established decisions, they can make errors that result in negative (legal) effects for individuals. When administrative procedures are developed by an algorithm, regardless of whether the system is only supposed to act in an automatic manner or has self-learning abilities, the fact that the decision made or implemented by the algorithm has legal effects on individuals means that the software programme must necessarily be considered and treated as a legal instrument.

1.2. ALGORITHMS ARE REGULATORY INSTRUMENTS

Once the legal effects of algorithms used in administrative procedures have been determined, it is also necessary to establish whether algorithms are administrative decisions, regulatory instruments or should be categorised as something altogether different, given that the controls and requirements they are subjected to will vary depending on the legal qualification they receive.

[1582] *Ibidem*.

When algorithms are designed, they are created as sets of instructions to be applied to specific cases. These instructions establish how to process the data that they are fed. Machine learning systems evolve, they learn from the data they process and continuously shift their parameters accordingly, meaning the set of instructions might change over time. However, this does not change the reality that the actual algorithm is a set of rules, an abstraction of reality to which data is fed in order to obtain resolutions for particular cases. Hence, the decision that results from the automated processing of data is the equivalent to an administrative decision while the actual algorithm constitutes the general parameters that must be followed and applied to each particular case, that is, a regulatory instrument. Therefore, while the result of algorithmic decision-making in specific cases can be characterised as an administrative decision, the algorithm that is used and its source code must be treated as a regulatory instrument.[1583] In addition, the automated processing phase substitutes the public servant that would generally make the decision on how a rule should be applied in a specific case.

1.2.1. Proposals that reject the regulatory nature of algorithms

Some legal scholars consider algorithms cannot be qualified as regulatory instruments. In this sense, ARROYO JIMÉNEZ has argued that algorithms should be qualified as soft law instruments and not as regulatory tools, for algorithms are not recognised as regulations by legal systems and therefore do not act as validating elements for ulterior administrative decisions. In other words, if an administrative decision is contrary to the rules set out by the algorithm said decision would not be nullified and eliminated, only if said decision infringed the provisions of a traditional regulatory instrument would the mechanisms to declare it null and void be activated.[1584] He concludes algorithms are soft law tools and can therefore not be categorised as regulatory instruments or administrative decisions.

This idea is not correctly approached for several reasons. The decisions (administrative resolutions) that result from algorithmic processing cannot contravene the rules of the algorithm. The phases of traditional administrative decision-making processes are clearly divided in three distinguishable parts in which (1) a written regulation is (2) applied by a

[1583] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, p. 236; YEUNG, K., "Algorithmic regulation…", *cit.*, 2018, p. 507.
[1584] ARROYO JIMÉNEZ, L., "Algoritmos y reglamentos", *Almacén de Derecho*, 25th February 2020. Available on 1st April 2020 at: https://almacendederecho.org/

public servant to the specific case, which results in (3) an administrative decision. It is therefore possible for the public servant to erroneously apply the regulatory instrument.

However, in the case of algorithmic processing, the general rules contained in the algorithm are applied to each specific scenario when the data for that case is introduced in the programme. Hence, the whole process is inextricably linked in a manner which precludes the possibility of arguing that the results provided by the algorithm were contrary to the algorithm's general rules. In cases in which the provisions of a traditional regulatory instrument are encoded into the algorithm, which then applies said provisions to each specific case, it might be argued that algorithm designers made mistakes when coding the traditional regulatory instrument or that the algorithm worked in a way it was not supposed to, but it is very difficult (if not impossible) to argue that the decision that resulted from algorithmic processing contravened the very rules that structure the automated system. If anything, a series of rules that established the relationship between the manually written and digitally coded regulatory instruments should be established in those cases in which the latter are based on the former. The relationship between both elements should be articulated through systems' functional specifications.

Moreover, this particular proposal has no practical translation, as its implications are limited to the theoretical level. On a practical level, ARROYO JIMÉNEZ does recognise the need for a series of guarantees that are similar if not the same to the ones applied to regulatory instruments.[1585] However, precisely in this point is were differences with the requirements and guarantees set with regard to regulatory instruments could be established. Algorithms do present a series of characteristics that hamper a comprehensive application of the requirements set for regulatory instruments. For example, the constant self-updating that some algorithms carry out, heavily hinders the possibility of publishing every modification to the programme, as it is demanded in the case of regulatory instruments.

PONCE SOLÉ[1586] also argues that, when algorithms are used in the exercise of administrative powers, they should not be treated as regulatory instruments, that is, sources of law, but as somehow equivalent to the human civil servants that previously made decisions that are now

[1585] ARROYO JIMÉNEZ, L., "Algoritmos y reglamentos", *cit.*, 2020.
[1586] PONCE SOLÉ, J., "Inteligencia artificial, Derecho administrativo y reserva de humanidad: algoritmos y procedimiento administrativo debido tecnológico", *Revista General de Derecho Administrativo*, No. 50, 2019, pp. 1-52.

delegated to machines.[1587] He does, however, recognise the need for algorithmic "transparency and citizen participation in their design",[1588] which include "the construction of an administrative due process for approving algorithms; the reassertion of a right to an understandable explanation of the way in which algorithms operate based on the right to understand; the use of open programming; active advertising and the right to access the algorithms; the prevention and control of bias in the algorithms and the data protection impact assessment of the algorithms".[1589]

There are several elements in this doctrinal proposal to which attention should be drawn. Firstly, it is not possible to open up the human mind and examine the real reasons why an individual makes a decision. Conversely, this is something that can be done, at least to a certain extent, when it comes to algorithmic decision-making. Algorithms might sometimes be opaque due to their complexity or their owner's willingness to keep them secret under the protections offered by intellectual property rights and trade secrets. It is however possible to examine many elements relevant to the system's design and experiment with the algorithm, providing a level of insight that cannot be achieved when decisions are made by humans.[1590] This quality offers the possibility of framing algorithmic control and accountability in a way that is inspired by existing control and review mechanisms applied to regulatory instruments. If the legal theory of algorithmic decision-making is constructed from a perspective that prioritises their similarities to human decision-makers instead of treating them as the non-sentient objects that they are, we risk losing sight of some of the real possibilities that exist for establishing *ex ante* and *ex post* controls on these types of instruments.

Secondly, while some of the algorithms used by the public sector are simple automatic systems, which are supposed to work in a straightforward manner (applying the rule when its factual assumption takes place), an increasing number of complex algorithms are being used. In the case of purely automatic simple systems, mistakes can be made when applying the rule, but using traditional administrative review mechanisms in combination with a few transparency requirements could solve these problems relatively easily. However, complex machine learning systems can change their initial parameters, that is, the rules that they are applying, and reach solutions by analysing a number of variables and drawing a series of

[1587] *Idem*, p. 35.
[1588] *Ibidem* (own translation).
[1589] *Idem*, pp. 36-37 (own translation).
[1590] KLEINBERG, J. *et al.*, "Discrimination in the age of algorithms", *cit.*, 2018, p. 114.

inferences that largely differ from those that are carried out by humans and from the written regulations that they are based on, meaning they do, in practical terms, act as a source of law.[1591]

In any case, the most important aspect of recognising the regulatory nature of automated systems is precisely providing citizens who are subjected to algorithmic decision-making with sufficient guarantees.[1592] Hence, what is especially relevant with regard to these proposals is recognising the legal nature of algorithmic systems used in public decision-making and the need to establish sufficient safeguards and guarantees.

1.2.2. Administrative court of Lazio-Roma, Judgment No. 3769

As it was already indicated, the Lazio-Roma Administrative court[1593] ruled that an algorithm used by the Italian Ministry of Education to manage teacher mobility was an administrative decision. However, the reason behind qualifying algorithms, or at least this specific algorithm, as administrative decisions in this particular ruling resulted from the discussion having focused on whether algorithms should be qualified as legal instruments at all.

Hence, in this particular case, the issue was far removed from the discussion regarding whether algorithms should be qualified as administrative decisions or regulatory instruments. However, when observing the court's reasoning, it is clear that it recognises that the programme contains a series of general instructions that it then proceeds to apply to specific cases:

> "By source language we mean the text of a programmed algorithm written in a programming language by a programmer in the programming phase; consequently, a computer programme is the expression of an organised and structured set of instructions contained in any form or medium capable, directly or indirectly, of causing a given function, task or result to be performed or obtained by means of an electronic information processing system…"

[1591] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, pp. 230-234.
[1592] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, pp. 249-261; DE LA CUEVA, J., "Código fuente, algoritmos y fuentes del derecho", *El notario del siglo XXI*, No. 77, 2018. Available on 18th June 2020 at: https://www.elnotario.es/
[1593] Administrative Regional Court of Lazio-Roma, Section III bis, Judgment No. 3769, 22nd March 2017.

> "… the processing of the administrative decision's content is entirely entrusted to the computer tool and therefore ultimately to the machine, which is directly responsible for the retrieval, connection and interrelationship between the rules and the data, thus assuming an instrumental role with regard to the final administrative act. In the case mentioned, it is the processing of the content of the act itself that is carried out electronically, processing that consists, precisely, in the logical process that leads to the drafting of the final act in relation to the respective content and that specifies its justification…"

The cited paragraphs must be interpreted together in that the algorithm contains the instructions (regulation) that it then applies to the particular case. Even though the Court determined that the algorithm is an administrative decision, it clearly recognised the specific elements that lead to the need of categorising algorithmic tools as regulatory instruments but did not discuss said issue, as it was not brought forward as a contentious matter.

1.2.3. Solely automated non-binding and semi-automated decision making

In some cases, automated systems are used in administrative procedures in which there is some form of human intervention in the decision-making process or in which algorithmic outputs are not binding.[1594] Public administrations can use risk-predictors and other types of automated systems in order to inform decisions or produce non-binding outcomes rather than as fully automated decision-making systems.[1595] The effect algorithms would have in these cases is similar to that of expert opinions requested in administrative procedures.

However, even when an algorithm is not to be used for providing binding decisions but only as a form of advisory or support tool, they must also be treated, to a certain extent, as regulatory instruments. In this sense, it is convenient to draw an analogy with the way in which the activity of administrative advisory bodies is regulated. Advisory bodies draw their reports and carry out their activity based on the legal instruments that regulate them. An algorithm used for similar purposes to the ones exercised by advisory bodies encompasses both, the rules upon which the opinion it issues are based (equivalent to the legal instruments that regulate advisory bodies) and the procedure leading to the opinion it issues (equivalent to the activities and procedures carried out by advisory bodies when issuing an opinion).

[1594] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, p. 237.
[1595] OSWALD, M., "Algorithm-assisted decision-making in the public sector…", *cit.*, 2018, pp. 2-3.

Consequently, although the effects the algorithm has in the procedure are not as relevant as when fully automated and binding decision-making is employed, they must also be treated as regulations to the extent they decide how the opinion is issued.

1.2.4. The importance of recognising the regulatory nature to algorithms

As it has already been indicated, several scholars have shown their scepticism with regard to the regulatory nature of algorithms. We must agree with them in that algorithms are not the same as traditional regulations. In fact, they differ significantly in many aspects. However, while discussions are being held on whether algorithms should be considered to be regulations, administrative decisions or something altogether different, these systems are being increasingly used in administrative procedures that affect the rights of individuals without being subjected to practically any control.

For example, the automated systems used by taxation agencies in Spain to flag individuals that are considered at risk of committing tax fraud are not made public.[1596] While this opacity may be justified in order to avoid gaming strategies, the fact is that the effects these algorithms may have on different fundamental rights, including the rights to equality and non-discrimination, are not being controlled.

Qualifying algorithms used in the public sector as regulations is thus not only useful due to the similarities of automated systems with regulatory instruments, but also necessary in order to draw attention to the risks generated by public algorithmic decision-making and the need to establish an effective framework of control for which we already have some of the tools in place. Theoretical discussions on the actual nature of public algorithms as instruments that should be encompassed under a new legal category, as interesting as they are, generate a vacuum that is currently being filled by the unprecedented and uncontrolled use of automated systems that make decisions that affect the daily lives of citizens. Applying the requirements set for regulatory instruments to algorithms used by the public sector may not be the best solution in the long-term, but in practical terms, is the only solution for the time being.

[1596] Resolution, 11th January 2019, of the General Directorate of the Spanish Tax Administration Agency, approving the general guidelines of the Annual Tax and Customs Control Plan for 2019.

1.3. THE PRINCIPLE OF LEGALITY MUST APPLY TO THE PUBLIC USE OF ALGORITHMS

A final consideration that must be drawn with regard to the use of algorithms by the public sector is the limits to which these systems must be subjected to; limits that are derived from the principle of legality. The underlying principle upon which private interactions are built is freedom of choice, autonomy and the notion of the negative binding nature of law: all that is not forbidden is allowed. Conversely, public institutions are generally positively bound by the content of the law: what is not allowed cannot be done. The positive binding nature of the law must, however, be nuanced, for limiting the possible actions of public administrations to what is exactly permitted by law would cause huge inefficiencies in their activities. Nonetheless, especially when it comes to public actions that limit citizens' freedom and those which require public spending, it is necessary for legal instruments to detail the actions that may be taken by public administrations to a sufficient degree. Moreover, any public action that may limit the fundamental rights and freedoms of individuals must necessarily be authorised by parliament and, in the case of Spain, by an organic Act, which requires a qualified majority of votes.

An especially worrying consequence that can be derived from the "SyRI" case is the fact that automated systems are being developed and used by public administrations without the proper legal authorisations. Once again, this is why qualifying algorithms as regulatory instruments is so important. The material perspective of the principle of legality limits the scope of regulation and actions of public powers, especially when it comes to the regulation or development of actions that affect especially sensitive elements, such as fundamental rights and freedoms. Public administrations cannot, under any circumstances, use their discretionary powers to carry out actions that may affect the fundamental rights and freedoms of individuals without the relevant and comprehensive legal authorisation.

1.4. FRICTIONS BETWEEN TRADITIONAL AND ALGORITHMIC REGULATION

An issue that must be considered both when regulating algorithms and when regulating by algorithm, is the difficulty of adapting traditional regulatory techniques to the nature of automated systems.[1597] The tendency towards ambiguity that characterises both regulatory and policy processes unavoidably clashes with the very detailed specifications that are

[1597] KROLL, J. A., "Accountable algorithms", *cit.*, 2017, pp. 695-696.

sometimes needed in algorithmic design[1598] in order to ensure that automated systems behave in a certain way.[1599]

This matter becomes especially relevant and visible when analysing some of the trade-offs that regulators will encounter and which will force very specific choices to be made when designing algorithms that were previously made on a case-by-case basis.[1600] For example, in the case of algorithms that predict recidivism, there will be a certain level of false positives that will have to be accepted when the algorithm is deployed.[1601] Depending on where this level is set, it could mean shifting some of the core principles of our justice system.[1602]

The principles agreed upon in the 2017 Asilomar Conference establish the need to set the general objectives that determine the kind of AI that society wants to aim for.[1603] These more general goals must, in turn, be transformed into more specific values and coded into algorithms.[1604] Hence, with regard to the main issue studied in this thesis, that is, the way in which algorithmic decision-making can affect the right to equality and non-discrimination, it will be necessary to agree on the level of equality that is deemed acceptable which may mean bringing about certain restrictions and requirements applicable to the private use of algorithms that would probably not be considered in a non-automated context.

Establishing these specific objectives that must be considered when designing algorithms is also essential given the role that programmers play when designing algorithms used by public bodies, especially when said algorithms are developed by private firms. For example, if public authorities do not establish the level of false positives and false negatives they deem acceptable in a tool for predicting tax fraud, the person in charge of creating the automated

---

[1598] The specific nature of the commands contained in automated systems must not be mistaken for a high degree of accuracy. These systems are representations of reality and in many cases work by making generalisations when they analyse the data that they are fed. This means that they treat individuals as members of groups and can therefore treat different situations as if they were the same. See BRAUNEIS, R. & GOODMAN, E. P., "Algorithmic transparency for the smart city", *cit.*, 2018, p. 123: "Government use of predictive algorithms poses an inherent challenge to traditional notions of fairness. By their nature, predictive models are simplifications that cannot consider all possible relevant facts about subjects, and that therefore necessarily treat people as members of groups, not as individuals. Generalizations, which may treat unlike cases alike, are inherent to this process".

[1599] KROLL, J. A. *et al.*, "Accountable algorithms", *cit.*, 2017, pp. 695-696.

[1600] KLEINBERG, J. *et al.*, "Discrimination in the age of algorithms", *cit.*, 2018, p. 119.

[1601] SCATAMBURLO, T., CHARLESWORTH, A. & CRISTIANINI, N., "Machine decisions and human consequences", *cit.*, 2019, pp. 62-69.

[1602] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, p. 233.

[1603] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, pp. 226-227; FUTURE OF LIFE, "Asilomar AI principles", 2017. Available on 14th March 2020 at: https://futureoflife.org/

[1604] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, pp. 253-254.

tool, who will probably work for a private firm, will be making these decisions, which are effective policy choices that individuals and bodies vested with public authority should be making.[1605] The humans behind the algorithms have a significant amount of discretionary power that must be limited and controlled by the competent authority.

2. CONSIDERATIONS REGARDING THE REGULATION OF ALGORITHMS EMPLOYED BY THE PRIVATE SECTOR

2.1. THE PRECAUTIONARY PRINCIPLE

Throughout the past few decades the regulation of economic activities has evolved towards a de-regulatory trend. In Europe, the Bolkenstein Directive was key in pushing this evolution forward and has been largely successful in its aim to advance the construction of the European single market.[1606] Said de-regulatory trend has especially taken place by reducing the administrative burden placed on citizens when initiating most economic activities. Prior administrative authorisations, which were previously required in order to kick-start many ventures, have been largely substituted in some economic sectors by other mechanisms, such as declarations of responsibility, which allow economic actors to initiate their activities in a practically automatic manner and delay administrative inspections to the moment after economic activities have started their operations.[1607]

While administrative simplification is convenient in order to enhance economic growth, it should not be pursued to such an expansive extent as the one witnessed in Europe without prior consideration of the implications it may have for other public interests.[1608] For instance, there are some economic activities that do not present significant risks for public interests or whose risks public administrations can easily control once they are operating. In that case, it does not make much sense to impose prior approval requirements which force economic actors to wait for public bodies to grant them the necessary authorisations in order to open up and develop their businesses. However, when dealing with activities that generate greater hazards[1609] and, in particular, when speaking of on-going technological developments and the

[1605] ZARSKY, T., "Transparent predictions", *cit.*, 2013, pp. 1518-1519.
[1606] NOGUEIRA, A., "La termita Bolkestein", *El Cronista del Estado Social y Democrático de Derecho*, 2011, pp. 58-59.
[1607] *Idem*, p. 60.
[1608] *Idem*, p. 70.
[1609] The words "risk" and "hazard" will be used indistinctly throughout this chapter.

uncertainty that comes with them, it is necessary to consider building the regulatory approach to the risks brought by algorithms through and based on the precautionary principle.[1610]

The origins of the precautionary principle can be traced back to the German regulation on environmental protection.[1611] Similarly, its adoption and evolution in the EU legal framework has also mostly taken place through environmental protection policy papers, regulatory provisions and instruments as it for instance appears in article 191.2 of the Treaty for the Functioning of the European Union, which establishes that "Union policy on the environment shall … be based on the precautionary principle". However, the possibility of applying said principle in other areas was fully acknowledged by the EU Commission, whose Communication on the precautionary principle indicated that:

> "Although the precautionary principle is not explicitly mentioned in the Treaty except in the environmental field, its scope is far wider and covers those specific circumstances where scientific evidence is insufficient, inconclusive or uncertain and there are indications through preliminary objective scientific evaluation that there are reasonable grounds for concern that the potentially dangerous effects on the environment, human, animal or plant health may be inconsistent with the chosen level of protection."[1612]

With regard to the interpretation of the precautionary principle, the Court of First Instance of the European Union has established that "where there is scientific uncertainty as to the existence or extent of risks to [the protected legal interest], the Community institutions may, by reason of the precautionary principle, take protective measures without having to wait until the reality and seriousness of those risks become fully apparent"[1613] and defines risk as "a function of the probability that use of a product or a procedure will adversely affect the interests safeguarded by the legal order".[1614] It is therefore not necessary to fully demonstrate the existence of the risks caused by the regulated product as it is sufficient that the risk

[1610] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, pp. 225-227; COTINO HUESO, L., "Riesgos e impactos del big data, la inteligencia artificial y la robótica: enfoques, modelos y principios de la respuesta del derecho", *Revista general de Derecho administrativo*, No. 50, 2019, pp. 27-28.
[1611] DOMÉNECH PASCUAL, G., *Derechos Fundamentales y Riesgos Tecnológicos*, *cit.*, 2006, pp. 255-256.
[1612] EU COMMISSION, "Communication from the Commission on the precautionary principle", COM/2000/0001 final, 1st February 2000, section 3.
[1613] Court of First Instance (Third Chamber) Judgment 11th September 2002, C-T-13/99, Pfizer Animal Health SA v. Council of the European Union, paragraph 139.
[1614] *Idem*, paragraph 147.

"appears nevertheless to be adequately backed up by the scientific data available at the time when the measure was taken".[1615]

As it has been proven throughout the course of this dissertation and has been widely documented by the scholarship, algorithms generate significant risks for the rights to equality and non-discrimination as well as for other fundamental rights and public interests. These risks, while known, appear within a framework of uncertainty regarding the possible future developments of algorithmic systems and the specific behaviour that each of said tools will develop, especially in the case of self-learning algorithms.[1616]

The proposal of applying the precautionary principle to the regulation of algorithms is thus not put forward in a void but attending to the context and nature of automated systems and, given the similarities of the risks they generate with those produced by certain activities on the environment, to the convenience of implementing similar regulatory mechanisms to those applied in environmental law, which will be discussed in the following section.[1617] In this sense it is interesting to highlight that, for instance, the Spanish Constitution while not explicitly mentioning it does indirectly foresee the application of the precautionary principle to data processing technologies[1618] when it establishes that "the law shall restrict the use of data processing in order to guarantee the honour and personal and family privacy of citizens and the full exercise of their rights" (article 18.4).

2.2. THE ENVIRONMENTAL POLLUTION ANALOGY

Several accounts have drawn analogies between the harms caused by data collection and processing technologies and the harms caused by environmental pollution, as well as the regulatory instruments used in both cases. The parallelisms between both phenomena are especially useful in order to justify public intervention in the private use of algorithms.

[1615] *Idem*, paragraph 144.
[1616] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, p. 227; MAZUR, J., "Automated decision-making and the precautionary principle in EU law", *cit.*, 2019, p. 7.
[1617] BEN-SHAHAR, O., "Data pollution", *cit.*, 2019, pp. 104-159; MAZUR, J., "Automated decision-making and the precautionary principle in EU law", *cit.*, 2019, pp. 3-18.
[1618] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, p. 239.

2.2.1. Environmental law mechanisms used in the European data protection framework

MAZUR shows that some of the instruments adopted in the GDPR imitate the techniques used in environmental regulation.[1619] Said similarities can mainly be drawn with regard to the rights to information[1620] and two of the collaborative governance tools contained in the Regulation: DPIAs[1621] and certification systems. However, while the incorporation of governance tools, in particular DPIAs, is unquestionably inspired by environmental law[1622] the differences between the way in which information rights are incorporated to data environmental protection law comes to show the shortcomings of privacy approaches, which do not fully grasp the public nature of algorithmic-generated hazards.[1623]

The GDPR and other privacy instruments include a set of rights to access and information. However, as it was already indicated, it builds said rights from the perspective of individual protection and does not complement them with solid general accountability mechanisms for public participation in the control of algorithms. This stands in direct contrast to the public participation inspired regime contained in environmental law, which can, for instance, be found in the EU's Directives on public access to environmental information[1624] and on public participation in the drawing up of certain plans and programmes relating to the environment.[1625]

Additionally, as it was already indicated in part I, the "best available techniques" mandate has, to a certain extent, been translated to the "data protection by design and by default" rule, for article 25 of the GDPR states that:

> "Taking into account the state of the art, the cost of implementation … the controller shall … implement appropriate technical and organisational measures, such as

[1619] MAZUR, J., "Automated decision-making and the precautionary principle in EU law", *cit.*, 2019, pp. 8-10.
[1620] *Idem*, pp. 8-9.
[1621] *Idem*, pp. 9-10.
[1622] BINNS, R., "Data protection impact assessments: a meta-regulatory approach", *International Data Privacy Law*, vol. 7, No. 1, 2017, p. 23; DE HERT, P., "Data protection as bundles of principles, general rights, concrete subjective rights and rules: piercing the veil of stability surrounding the principles of data protection", *European Data Protection Law Review*, vol. 3, No. 2, 2017, p. 174.
[1623] BEN-SHAHAR, O., "Data pollution", *cit.*, 2019, p. 118-131.
[1624] Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to environmental information and repealing Council Directive 90/313/EEC.
[1625] Directive 2003/35/EC of the European Parliament and of the Council of 26 May 2003 providing for public participation in respect of the drawing up of certain plans and programmes relating to the environment and amending with regard to public participation and access to justice Council Directives 85/337/EEC and 96/61/EC - statement by the Commission.

pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects".

2.2.2. Similarities between the harms caused by environmental pollution and data processing technologies

With regard to the parallelisms that can be drawn between the harm and protection of the natural and digital environments BEN-SHAHAR points out to three elements: the effects of environmental and data pollution on public interests, other people and "precaution and insurance externalities"[1626] which, in turn, fall under the overarching idea that the harms caused by data are public and should therefore be addressed and prevented through a regulatory system built from a public law perspective.[1627]

In particular, with regard to algorithmic discrimination, he indicates that there are many benefits to the use of automated data processing systems for they can help produce more accurate and personalised results in a series of different areas in which humans interact with algorithms, such as medicine or education. However, said benefits are generally internalised by tech companies and other actors using algorithms while "negative externalities, in contrast, remain orphan. They affect groups too broad and dispersed and cause injuries that are too abstract for private remedies to be effective" and cause the general degradation of the public digital environment.[1628]

2.3. MARKET FAILURES AND OTHER PROBLEMS GENERATED BY THE DATA SERVICES SECTOR

The harms and risks that the use of algorithms has proven to cause to the fundamental rights and freedoms of individuals and to other public interests and the failure of private market powers to control and deal with said harms is cause for regulatory intervention in the data services sector. Nonetheless, the following pages briefly address the market failures of the data services sector in order to emphasise the need for public intervention.

[1626] BEN-SHAHAR, O., "Data pollution", *cit.*, 2019, pp. 112-118.
[1627] *Idem*, pp. 131-148.
[1628] *Idem*, p. 115.

2.3.1. Negative externalities

The use of data collection and processing technologies degrades the social environment by generating harms to certain public interests and shared values, including equality and non-discrimination,[1629] both as subjective rights and general principles that Western democracies and, in particular, EU member states must abide by. Numerous examples of algorithmic discrimination and of the perpetuation of structures of inequality through the use of algorithms have been conveyed throughout the dissertation. Examples of algorithms that discriminate or perpetuate structures of discrimination by making errors that especially affect members of disadvantaged groups are widespread and range from instances of stereotyping in advertising[1630] to limiting the access of protected groups to goods and services[1631] and employment[1632] or reinforcing the structural discrimination present in law enforcement against certain racial minorities.[1633]

A particularly relevant issue regarding the discriminatory results that algorithms can yield is that, in many cases, these do not directly affect many of the individuals that have provided their personal data. Since algorithms generate predictions that are based on the aggregate data of many individuals, the actual effects of algorithmic decision-making will, in many cases, not affect all the individuals whose data has been fed to the system.

There are also many other externalities, such as inadvertently sharing third-parties' data with social platforms[1634] or the spreading of fake news that aim to manipulate individuals and affects democratic processes. The negative externalities caused by these technologies do not exclude the existence of many positive effects. It is, however, important to acknowledge and address the existence of these issues.

[1629] *Idem*, pp. 114-115.
[1630] DATTA, A., TSCHANTZ, M. C. & DATTA, A., "Automated experiments on ad privacy settings…", *cit.*, 2015, pp. 92-112; SWEENEY, L., "Discrimination in online ad delivery", *cit.*, 2013, pp. 44-54.
[1631] ANGWIN, J. & PARRIS JR., T., "Facebook lets advertisers exclude users by race", *cit.*, 2016.
[1632] ALLHUTTER, D. *et al.*, "Algorithmic profiling of job seekers in Austria…", *cit.*, 2020, pp. 1-17; DASTIN, J., "Amazon scraps secret AI recruiting tool that showed bias against women", *Reuters,* 10th October 2018. Available on 12th February 2019 at: https://www.reuters.com/
[1633] O'NEIL, C., *Weapons of Math Destruction…*, *cit.*, 2017, pp. 84-104.
[1634] COASTINE, J., "Facebook is shutting down its API for giving your friends' data to apps", *TechCrunch*, 28th April 2015. Available on 13th June 2020 at: https://techcrunch.com/

2.3.2. Monopolistic behaviour

The implications of categorising algorithms used by the public sector as regulatory instruments are quite significant seeing as it does mean accepting a paradigm shift from written to technology-based law and, as was already indicated, there are relevant differences between the way in which the creation and implementation of traditional regulatory code has worked and the way in which algorithmic regulatory code works.[1635] However, the large power that governments have traditionally held means that Western democratic states have developed over the past few centuries a system of guarantees and accountability that can be adapted and applied to algorithms, once their regulatory nature is fully recognised. Hence, regulating and controlling the use of algorithms by the public sector will, in theory, not pose as many problems as regulating the private use of algorithms.

Western societies have developed upon the idea that governmental actions mean control while the development of private activities is an expression of freedom.[1636] The degree to which this notion of the public/private sector dichotomy as regulation vs. freedom is held, varies depending on the political culture of each country. However, the idea that private parties have a right to act and develop their activities in a free manner and that governments can only limit their actions if certain risks are generated is commonly accepted. Consequently, the very idea that private sector organisations may act as regulators seems counterintuitive. Thus, suggesting that similar constraints to the ones placed on the public sector should be placed on certain private activities may be difficult to accept. However, the power acquired by certain private organisations that currently own a large amount of capital and, more importantly, information, means that said organisations behave as quasi-regulators.[1637]

YEUNG[1638] follows from BLACK's[1639] definition of regulation in order to conclude that algorithms, whether used by public or private bodies, are regulatory instruments. YEUNG defines algorithmic regulation as "decisionmaking systems that regulate a domain of activity

[1635] LESSIG, L., *Code version 2.0*, *cit.*, 2006, pp. 5-6.
[1636] *Ibidem.*
[1637] KIM, N. S. & TELMAN, D. A., "Internet giants as quasi-governmental actors and the limits of contractual consent", *Missouri Law Review*, vol. 80, No. 3, 2015, pp. 723-770.
[1638] YEUNG, K., "Algorithmic regulation…", *cit.*, 2018, p. 507.
[1639] BLACK, J., "Learning from regulatory disasters", *LSE Law, Society and Economy Working Papers* 24/2014, 2014, p. 3. Available on 6th March 2020 at: http://eprints.lse.ac.uk/: "regulation or regulatory governance is the organised attempt to manage risks or behaviour in order to achieve a publicly stated objective or set of objectives".

in order to manage risk or alter behavior through continual computational generation of knowledge from data emitted and directly collected (in real time on a continuous basis) from numerous dynamic components pertaining to the regulated environment in order to identify and, if necessary, automatically refine (or prompt refinement of) the system's operations to attain a prespecified goal."[1640] By adopting this broad concept of regulation it becomes much easier to see how the private use of algorithms is, in fact, a form of regulation. It does however fall short in justifying why we should equate the public and private use of algorithmic regulation given that the development of some forms of regulation by private firms is not a new phenomenon. After all, every organisation needs a minimum set of rules.

It is therefore necessary to combine said concept of regulation with the reality of the space (cyberspace)[1641] in which private algorithmic regulators operate and the power they hold. By assessing the real power held by some private organisations, it is possible to conclude that, since many of the private firms creating algorithms and controlling much of the world's data behave as monopolies, they have, to a certain extent and in certain contexts, acquired regulatory powers that are not far from the regulatory powers held by public institutions. Companies such as Alphabet (Google), Facebook, Apple or Amazon, occupy a very important segment of online markets and are the main collectors of individuals' personal data, which they then use for their own benefit or sell to third parties (including governments). For example, Google is the main search engine and online advertising services provider, and has, in fact, been fined by the European Commission for abusive practices in online advertising.[1642] In addition, all four of the aforementioned companies are currently undergoing antitrust investigations both in the US and the EU.

However, for the time being, the instruments contained in antitrust regulations have failed to reduce the monopolistic power of these companies. In this context, the recently published majority staff report and recommendations on the investigation carried out by the US Congress' Subcommittee on antitrust, commercial and administrative law of the Committee on the judiciary, offers very illustrative insights on the need to limit the monopolistic power

[1640] YEUNG, K., "Algorithmic regulation…", *cit.*, 2018, p. 507.
[1641] LESSIG, L., *Code version 2.0*, *cit.*, 2006.
[1642] EUROPEAN COMMISSION, "Antitrust: Commission fines Google €1.49 billion for abusive practices in online advertising", 20th March 2019. Available on 20th June 2020 at: https://ec.europa.eu/

of Alphabet (Google), Facebook, Apple and Amazon and to strengthen and revive antitrust laws.[1643]

Moreover, antitrust instruments are not aimed towards protecting fundamental rights, such as the rights to equality and non-discrimination, and must therefore be complemented by other regulatory instruments aimed towards comprehensively controlling and, to the extent that is possible, preventing the harms caused by the algorithms developed by these firms.

These firms operate spaces that have become crucial in the daily lives of many individuals and, as owners of said spaces, regulate the way in which many interactions are conducted within them and can make decisions that can have a very significant impact on the humans that directly or indirectly interact with the machine.[1644] Furthermore, the control these firms exercise over their sector of cyberspace means that they create choice architectures that limit and heavily influence individuals' autonomy in deciding how to behave.[1645] The extent to which algorithms can be used to limit the choices of individuals, whether it is in a direct or indirect manner, must be considered as part of the potential for regulation, as it was traditionally understood to be carried out by governments and public bodies, that private parties have through the use of automated mechanisms.

Moreover, regimes such as the one set up by the GDPR (and the Data Protection Directive before it), are delegating powers to private firms, which increases the power of the most powerful actors in the digital market even further.[1646] In this sense, a particularly prominent case of delegation is the one that has taken place with regard to Google and the right to erasure, commonly known as "the right to be forgotten" recognised in article 17 of the GDPR.[1647] Google is the one that requests for erasure are sent to, who analyses the cases and the applicability of the CJEU case law, [1648] who gathers and carries out any investigation that might be necessary and who enforces the right to be forgotten.[1649] It is important to highlight

[1643] NADLER, J. & CICILLINE, D. N., "Investigation of competition in digital markets", Subcommittee on antitrust, commercial and administrative law of the Committee on the judiciary, October 2020.
[1644] EUBANKS, V., *Automating Inequality…*, *cit.*, 2017; NOBLE, S. U., *Algorithms of Oppression…, cit.*, 2018; O'NEIL, C., *Weapons of Math Destruction…*, *cit.*, 2017.
[1645] YEUNG K., "'Hypernudge'…", *cit.*, 2017, pp. 118-136.
[1646] KAMINSKI, M., "Binary governance…", *cit.*, 2019b, pp. 1529-1616.
[1647] LEE, E., "Recognizing rights in real time: the role of Google in the EU right to be forgotten", *University of California Davis Law Review*, vol. 49, No. 3, 2016, pp. 1017-1095.
[1648] CJEU Judgment 23rd May 2014, C-131/12, Google Spain SL and Google Inc. v. Agencia Española de Protección de Datos (AEPD) and Mario Costeja González; CJEU Judgment 24th September 2019, C-507/17, Google LLC, successor in law to Google Inc. v. Commission nationale de l'informatique et des libertés (CNIL).
[1649] LEE, E., "Recognizing rights in real time…", *cit.*, 2016, pp. 1066-1072.

that delegating certain control powers to private actors is not a new phenomenon nor is it necessarily negative or indicative of monopolistic power. It is, however, undeniable that the power held by these firms is further enhanced through the exercise of these delegated powers.

Private organisations that, like Google, are effective monopolies in their market segment should therefore be intervened by public institutions. The use of automated decision-making by these firms should be subjected to similar requirements as the use of algorithms by the public sector and the guarantees and protections recognised to citizens against this form of algorithmic regulation should also be drawn and structured in a similar manner to the ones granted when private persons interact with public bodies.

2.3.3. Asymmetric information and imperfect rationality

Although the issues that arise from "the privacy paradox" have already been addressed, it is relevant to briefly point them out. Firstly, individuals are not aware of the kind of information that they are actually providing when interacting in digital spaces or the purposes for which it can be used. The information collected on individuals is used to heavily manipulate them and make decisions that can directly impact the most significant parts of their lives. Furthermore, individuals are not aware of the extent to which these technologies affect them and perceive the risks of data collection and processing systems to be small and indirect.

Secondly, even though individuals do express privacy preferences, they then do not act accordingly and share their data. These actions are sometimes the result of lack of alternatives to the digital services individuals need to use. However, in other instances, it is the hyperbolic discounting of risks that characterises humans' evaluation of short-term gains against long-term risks, that leads individuals to prioritise the short-term benefits of accessing a digital service over the long-term risks or harms that the data they provide in exchange for accessing said service may cause them. The harms caused by data processing technologies are disperse and do not occur immediately after data is shared. Moreover, in many cases, the harms caused to an individual are the result of data shared by others. All in all, it is very difficult to entrust individuals with protecting themselves from harms they do not even perceive to be real.

2.3.4. Uncertainty

Uncertainty over the harms caused by the use of data collection and processing technologies is closely associated with the non-rational behaviours of consumers in digital environments.[1650] However, the notion of uncertainty and its implications in the regulation of algorithmic systems go beyond the personal knowledge on the risks that these technologies have for an individual and enter the realm of uncertainty over the social harms that they can cause. The extent to which the increasing use of automated systems will impact the social fabric of our societies is still unknown. However, the developments that have taken place over the last few years suggest that, while algorithms produce important benefits in certain areas, significant harms are still to be derived from their use if it not properly controlled.

3. GENERAL CONSIDERATIONS REGARDING ALGORITHMIC TRANSPARENCY

Rendering automated systems transparent (or at least understandable) is essential in order to detect instances of discrimination. Moreover, transparency is a relevant value on its own that is necessary to uphold in democratic countries, particularly when individuals interact with public bodies and especially when their fundamental rights are at stake. Transparency is in many cases necessary for accountability and citizen participation in public activities, consequently, most legal systems recognise the right to transparency and access to public information. While the transparency requirements for private bodies are not as demanding as those set for public institutions, when private decisions affect individuals' rights, particularly in cases of discrimination against traditionally disadvantaged groups in areas such as employment or access to goods and services, private parties must justify the decisions made, which indirectly grants individuals with some form of a right to transparency.

Much of the literature on regulating automated systems has focused on rendering algorithmic systems transparent either from an individual perspective, for example, through the right to explanation,[1651] or from the perspective of full system transparency.[1652] However, for many, the main focus has been set on making systems only transparent enough so that they can be

[1650] BEN-SHAHAR, O., "Data pollution", *cit.*, 2019, pp. 112-118.
[1651] WACHTER, S., MITTELSTADT, B. D. & RUSSELL, C., "Counterfactual explanations without opening the black box…", *cit.*, 2018, pp. 841-887.
[1652] KROLL, J. *et al.*, "Accountable algorithms", *cit.*, 2017, pp. 633-705.

sufficiently accountable without opening the black box.[1653] As with the other decisions that regulators have to make with regard to algorithms, transparency faces a trade-off with innovation, intellectual property and public security to name a few elements. Hence, the scholarship has aimed to develop tools through which to render algorithms accountable but without disclosing the actual source code.

As the previous chapter indicated, transparency on its own is not effective, which means that other mechanisms for rendering automated systems accountable must also be developed. However, part of the literature has become complacent with some of the arguments that are provided against full algorithmic individual and systemic transparency. Consequently, while not ignoring that transparency is not an absolute remedy for lack of algorithmic accountability, its still necessary to consider it as a necessary objective that aligns with the mandates of democratic states.

Some of the justifications for not making systems fully transparent refer to the inherent limitations of fully disclosing the way in which automated systems work or reach a particular decision. Providing full transparency may be technically challenging given the complexity of algorithmic systems, for it is very difficult to translate said systems in order for humans to understand them, especially when they employ machine learning technologies which provide algorithms with self-developing abilities.[1654] In addition, even when algorithms can be more or less explained it is still possible that the explanation provided is not understood by recipients and, when understandable explanations are available, they might still be not be intuitive.[1655]

Both these arguments are valid and refer to the reality of machine learning systems which are simply too difficult for the average citizen to understand. Strategies designed to provide individual explanations and systemic accountability without providing the source code,[1656] are necessary especially in order to ensure that individuals are provided with a justification

[1653] For example, see the proposals made by KROLL, J. *et al.* in "Accountable algorithms", *cit.*, 2017, pp. 633-705; BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, pp. 1085-1139; WACHTER, S., MITTELSTADT, B. D. & RUSSELL, C., in "Counterfactual explanations without opening the black box…", *cit.*, 2018, pp. 841-887; and, EDWARDS, L. & VEALE, M., in "Slave to the algorithm?…", *cit.*, 2017, pp. 19-84.
[1654] ZARSKY, T., "Transparent predictions", *cit.*, 2013, pp. 1519-1520.
[1655] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, pp. 1094-1199.
[1656] KROLL, J. *et al.*, "Accountable algorithms", *cit.*, 2017, pp. 633-705; WACHTER, S., MITTELSTADT, B. D. & RUSSELL, C., "Counterfactual explanations without opening the black box…", *cit.*, 2018, pp. 841-887; EDWARDS, L. & VEALE, M., in "Slave to the algorithm?…", *cit.*, 2017, pp. 55-60.

for the decisions that affect them but also so that the general public can understand algorithms, thereby increasing system legitimacy and, more importantly, enhancing third party oversight. However, this does not mean that other forms of transparency should not be provided but that the algorithmic transparency mandate should include both understandable explanations and full system (source code) transparency.

There are also extrinsic limitations to algorithmic transparency, which refer to the elements that enter into conflict with fully disclosing automated systems. Full transparency can hamper innovation and harm organisations developing said systems as source codes are proprietary information that, if released to the public, could be copied by other competing firms.[1657] Full transparency might sometimes enter into conflict with fundamental rights as it could require revealing personal data, especially when individual explanations are being provided.[1658] Finally, full system disclosure can also lead to gaming strategies, which is a problem that has especially been brought up with regard to the use of algorithms in law enforcement.[1659] Gaming, which has mainly been discussed in relation to public security, is also relevant in private service provision seeing as, if individuals are aware of the data points that for example an algorithm used by an insurance company correlates with "low-risk clients", they will be able to modify their behaviour and lifestyle to fall into the category that most benefits them.[1660]

The algorithms used by private organisations are, and should be, protected by intellectual property and trade secrets. There is an element that should nonetheless be noted to this regard, which is that, if rendering automated systems fully transparent does not actually help to understand them, why is it that private firms insist on keeping their information secret? This contradiction proves that, while average citizens may not be able to understand the algorithm, it is possible to effectively enact some form of oversight when source codes are disclosed. In any case, returning to the need to protect proprietary systems used by private organisations, there are ways in which to provide full disclosure of the algorithm without making said disclosure completely public, for example, by providing all information to a public institution. In this sense, BAROCAS and SELBST suggest establishing certain cases in

[1657] WACHTER, S., MITTELSTADT, B. D. & RUSSELL, C., "Counterfactual explanations without opening the black box…" *cit.*, 2018, p. 843.
[1658] *Idem*, p. 844.
[1659] BAYAMLIOĞLU, E., "Transparency of automated decisions in the GDPR…", *cit.*, 2018, p. 18; ZARSKY, T., "Transparent predictions", *cit.*, 2013, pp. 1553-1554.
[1660] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, pp. 264-265.

which full disclosure should be triggered, for example, upon litigation.[1661] This strategy would however have to be very carefully crafted in order to ensure that incentives are not created for competitors to draw false lawsuits in order to get access to source codes. In addition, a list of particularly risky algorithms that should be made fully transparent from the moment in which they are deployed could also be established. The need to protect individuals and groups from the risks posed by algorithms on their fundamental rights and freedoms cannot be overcome by companies' rights to keep proprietary information secret.

The risk of revealing personal data when providing explanations or full system transparency can easily be dealt with since in some cases it will not be necessary to provide information on data subjects. Moreover, even when information on individuals needs to be made available in order, for example, to analyse how a decision made regarding an individual compares to the same decision-making process affecting other individuals', it will be possible to offer a sufficient degree of anonymisation for the data points that provide information leading back to the individual or simply provide the data that is relevant to the decision. However, this may not even be a problem in cases in which providing aggregate data is sufficient in order to detect instances of discrimination.[1662]

As for the possibility that full algorithmic transparency may lead to gaming strategies, which is one of the main arguments wielded for keeping public algorithms secret, the literature has noted that the probability of gaming behaviour actually taking place is lower than it is generally thought.[1663] Moreover, the increasing complexity of automated systems significantly reduces the chances that individuals may have to game algorithmic decision-making systems.

On the one hand, machine learning systems are very difficult to understand at user-level. Even if their source code is published only specialised third parties will be able to disentangle their meaning and acquire knowledge on the shift in behaviour that would be required to yield different results when classifying individuals. Paradoxically, this could lead to greater

[1661] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, p. 1133.
[1662] HACKER, P., "Teaching fairness to artificial intelligence…", *cit.*, 2018, pp. 1173-1174.
[1663] EDWARDS, L. & VEALE, M., in "Slave to the algorithm?…", *cit.*, 2017, p. 63.

inequalities since only those with more resources would have the capacity to access the meaning of algorithms.[1664]

On the other hand, the amount of data that is currently available on individuals makes it very hard to mislead the machine because, even if an individual suddenly changes her behaviour, said drastic change will probably be detected and recorded. Furthermore, there are certain static characteristics that individuals cannot change and which will sometimes be the main variables used by algorithms in determining outcomes. In these cases, the possibility of gaming is practically out of the picture.

However, these concerns are valid and must be addressed in both the public and private sectors. As with many other issues that arise from the use of algorithms, gaming is not new. For example, taxation agencies have traditionally had to include strategies to prevent gaming when inspection plans were carried out.[1665] This may sometimes mean not offering full transparency but does ensure that a careful analysis is carried out in which different rights and interests are balanced out in order to determine how to combine the transparency requirements and rights necessary in a democratic state with other public interests that institutions must protect and realise.[1666]

Finally, it is important to be aware that the existence of conflicts between transparency and other interests cannot always serve as an argument in order to balance out competing elements. There are cases in which the aforementioned risks that might come with transparency simply do not operate due to the huge risks that the use of automated or semi-automated systems generate for the fundamental rights and freedoms of individuals. Particularly with regard to the risk of discriminatory decisions being made, it is important to highlight that the use of these systems offers the unprecedented possibilities in the prevention of discrimination. Existing legal frameworks force private parties to justify discriminatory decisions. However, as it will was discussed in part I of the dissertation, this mechanism only operates in an *ex post* manner. The use of algorithms in decisions that generate risks for the right to equality and non-discrimination of traditionally disadvantaged groups provides the possibility of carrying out *ex ante* and on-going controls to ensure that these systems are not

[1664] ZARSKY, T., "The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making", *Science, Technology, and Human Values*, vol. 41, No. 1, 2016, pp. 125-126.
[1665] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, p. 265.
[1666] *Ibidem*.

discriminatory. However, in order to do so it will be necessary to come to terms with the idea that, in certain cases, non-discrimination achieved through transparency must weigh more than the protection of a strictly economic definition of efficiency.

4. A SYSTEM OF PUBLIC INTERVENTION TO CONTROL ALGORITHMS

This section is based on some of the proposals that have been put forward by the scholarship in order to put forward a proposal of what could be a comprehensive system that ensures the control and accountability of algorithmic systems.

4.1. ORGANISATIONAL OPTIONS

For the system of algorithmic accountability to meet a minimum level of effectiveness it is necessary to provide either a single or several public bodies with the authority to carry out the regulatory and policy actions designed to control algorithms in general and, in particular, to prevent and deal with instances of algorithmic discrimination.[1667] There are different approaches that can be adopted in order to achieve public oversight, from increasing the remit of bodies and institutions in charge of controlling products that use algorithms so that they can also exercise oversight powers over the specific automated systems that they use, to setting up an independent agency specialised in algorithmic oversight.[1668]

4.1.1. Algorithmic control mainstreaming

The first option, providing institutions that control other areas with oversight powers over the automated systems used in the products they normally exercise control over, would help adapt the requirements set for each type of algorithm to the specific risks generated by the industry in which they are used. Moreover, it would reduce bureaucratic burdens and public expenditure since the increased costs derived from algorithmic oversight would not entail creating new agencies or bodies but simply providing certain institutions with a few more resources when necessary to carry out algorithmic oversight. Processors and controllers would also not have the obligation of undergoing several administrative procedures in

[1667] COUNCIL OF EUROPE COMMISSIONER FOR HUMAN RIGHTS, "Unboxing Artificial Intelligence: 10 steps to protect Human Rights", Strasbourg, Council of Europe, 2019, pp. 10-11.
[1668] HOFFMANN-RIEM, W., *Big Data. Desafíos también para el Derecho*, translation by Eduardo Knörr Argote, Cizur Menor, Aranzadi, 2018, pp. 154-157; TUTT, A., "An FDA for algorithms", *Administrative Law Review*, vol. 69, No. 1, 2017, pp. 83-123.

different institutions to have their products approved, as it would be sufficient to present additional documentation in the procedures that they are already used to undergoing.[1669]

However, this system presents significant shortcomings in that the control of algorithms would be very disperse and significantly diverge from one industry to another. This could lead to different requirements being set for the same type of algorithm used for two different purposes that generated similar levels of risks for the rights of individuals.[1670] Moreover, this system would fail to provide oversight for algorithms that are not integrated in products or processes that are already reviewed by public administrations and, considering the deregulatory trend that has taken place in Europe over the past decades, it is likely that only a reduced number of products that use algorithms which generate risks for the rights of individuals and for public interests, are subjected to public control.

More importantly, entrusting public bodies specialised in different areas of economic activity with, amongst other things, ensuring that algorithms are "fair", is not effective. This strategy would be the equivalent of equality mainstreaming in general policy, which, in many cases fails to produce the desired results because constructing policy and regulation from an inclusive perspective instead of building them based on dominant narrative values requires a degree of specialised knowledge in the existence of structural discrimination and its effects. Adding another level of oversight provided by equality or fundamental rights protection bodies could complement this system. Said bodies could help each of the public institutions overseeing algorithms to introduce fairness-related criteria when evaluating automated systems. However, existing equality bodies, at the European, national and regional/local level generally lack the means and expertise to help and intervene in the actions of other supervisory and regulatory bodies, especially when it comes to such a complex matter as automated decision-making systems.

4.1.2. Creating a non-independent supervisory task force or body

A second option would be to establish a non-independent specific task force or whole body, such as a ministry, in charge of supervising algorithmic systems.[1671] This was the strategy

[1669] TUTT, A., "An FDA for algorithms", *cit.*, 2017, p. 114.
[1670] *Ibidem.*
[1671] DJEFFAL, C., "Deutschland braucht nicht ein Digitalministerium, sondern viele!", *Süddeutsche Zeitung*, 18th September 2017. Available on 21st May 2020 at: https://www.sueddeutsche.de/; HOFFMANN-RIEM, W., *Big Data. Desafíos también para el Derecho*, *cit.*, 2018, p. 155.

adopted by the New York City Council, which, in 2018, passed a bill establishing the New York City Automated Decision System Task Force,[1672] which was supposed to analyse and detect instances of bias and discrimination in the public use of algorithms. This experiment largely failed due to the fact that public agencies were not mandated to provide the Task Force with the algorithms they used even when requested. This problem could however be easily solved by forcing public bodies to provide information on their algorithmic systems either on a regular basis or whenever it was requested. Nonetheless, taking up this option, whether it is in the form of a specific ministry or task force, is not advisable due to its lack of independence with regard to other public bodies, which will unavoidably reduce its effectiveness in exercising supervisory powers over the algorithms deployed by public administrations.

4.1.3. An independent supervisory agency

The final option is establishing independent agencies that supervise public and private algorithms. While several US scholars have put forward proposals for such an agency,[1673] which can provide very useful blueprints for European algorithmic oversight agencies, it is also necessary to be aware that the US system of public agencies is not equivalent to European independent agencies. The former have some level of independence in their supervisory powers, but mainly act as a ministerial branches that execute the US Federal Government's policy, largely due to the fact that it is the President of the United States who appoints the heads of said agencies. Conversely, European independent agencies or authorities, inspired by the German model, are articulated more as oversight bodies than as policymakers or implementers and therefore develop their activity in a significantly more independent manner.

The idea of setting up a European Agency for Artificial Intelligence and robotics was already put forward by the EU Parliament in its Resolution of 16th February 2017[1674] and further

[1672] RICHARDSON, R. (ed.), "Confronting black boxes: a shadow report of the new york city automated decision system task force," AI Now Institute, 2019.
[1673] SCHERER, M., "Regulating artificial intelligence systems: risks, challenges, competencies, and strategies", *Harvard Law & Technology Journal*, vol. 29, No. 2, 2016, pp. 353-400; SHNEIDERMAN, B., "Algorithmic Accountability: Designing for safety through human-centered independent oversight", *Turing Lecture (The Alan Turing Institute)* 31st March 2017. Available on 23rd May 2020 at: https://www.youtube.com/; TUTT, A., "An FDA for algorithms", *cit.*, 2017, pp. 83-123.
[1674] EUROPEAN PARLIAMENT, "European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics", *cit.*, 2017, recitals 15-17.

developed in the EU Commission's White Paper on Artificial Intelligence,[1675] which also suggests establishing a network of national agencies aimed towards facing the risks posed by new technological developments. This model makes sense within the European framework, in which much of the EU administrative structure has been articulated through these independent agencies with regulatory or semi-regulatory powers.[1676]

The current structure of Data Protection Agencies should be used in order to avoid contributing to the ineffective and relentless growth of public bodies. These agencies should expand their scope of activity to cover the protection of other fundamental rights and not just the right to data protection. In order to make these agencies effective, they should be integrated, first and foremost by experts in fundamental rights, in particular, on the protection of equality and non-discrimination, but also by technical experts focused on studying algorithmic fairness. These agencies should be invested with oversight and control powers as well as with the power to impose sanctions in order to effectively implement the system of algorithmic control that was previously put forward.

4.2. RISK-BASED MARKET APPROVAL OF ALGORITHMS

Risk-based approaches to regulation have been particularly prominent in the Anglo-Saxon world for quite some time but until more recently, have not started to be as successful in being explicitly adopted throughout continental Europe.[1677] This push towards risk-based regulation has been especially carried forward by the actions of European institutions, which have instilled this perspective on a number of regulatory sectors, including data protection.[1678]

[1675] EU COMMISSION, "White Paper on Artificial Intelligence – A European approach to excellence and trust", COM(2020) 65 final, 2nd February 2020, pp. 24-25.
[1676] VELASCO RICO, C. I., "Vigilando al algoritmo. Propuestas organizativas para garantizar la transparencia", in PUENTES COCIÑA, B. & QUINTIÁ PASTRANA, A., (dirs.), *El Derecho ante la Transformación Digital*, Barcelona, Atelier, 2019, pp. 81-82.
[1677] PAUL, R. & HUBER, M., "Risk-based regulation in continental Europe? Explaining the corporatist turn to risk in German work safety policies", *European Policy Analysis*, vol. 1, No. 2, 2015, pp. 5-33; ROTHSTEIN, H. *et al.*, "Varieties of risk regulation in Europe: coordination, complementarity and occupational safety in capitalist welfare states", *Socio-Economic Review*, vol. 17, No. 4, 2017, pp. 993-1020.
[1678] GELLERT, R., "Understanding the notion of risk in the General Data Protection Regulation", *Computer Law & Security Review*, vol. 34, No. 2, 2018, pp. 279-288; GONÇALVES, M. E., "The risk-based approach under the new EU data protection regulation: a critical perspective", *Journal of Risk Research*, vol. 23, No. 2, 2020, pp. 139-152.

Regulatory approaches based on risk measurement mainly serve two objectives: providing efficiency and institutional legitimacy.[1679] The latter is accomplished through the "*ex ante* determination of risk through assessment frameworks", [1680] which provides regulated entities knowledge regarding the way in which their activities will be controlled by public bodies and thus with the capability of developing the necessary mechanisms adapted to ensure each of their activities will pass all regulatory requirements. The efficiency and efficacy gains of risk-based regulation are twofold. On the one hand, they ensure that public bodies direct their resources mindfully, ensuring they are allocated proportionately to the hazards that the development of certain activities poses, therefore reducing unnecessary expenditures and exerting better oversight over especially risky activities. On the other hand, by targeting administrative control powers and requirements in a way that is proportionate to the hazards generated by each activity, administrative burdens are reduced whenever no significant risks are produced.[1681]

This approach to the regulation of algorithms also helps to justify the implementation of the precautionary principle. Basing the regulation of algorithms on said principle allows adopting especially restrictive (preventative) measures when the risks that algorithms pose on societal values and goods are particularly high and also justifies establishing specific public agencies to deal with these new technological developments. However, by combining the precautionary principle with a risk-based approach to the regulation of algorithms it is also possible to develop an efficient regulatory framework that does not hinder technological development by ensuring that the innovation principle is prioritised over precaution whenever no significant risks to the core values of democratic states are posed.[1682]

Risk-based approaches to the regulation of algorithms have been put forward, for instance, in the European Commission's White Paper on Artificial Intelligence[1683] and the German Data Ethics Commission's opinion on algorithmic systems[1684] and has also been suggested in other

[1679] BLACK, J., "The emergence of risk-based regulation and the new public risk management in the United Kingdom", *Public Law*, No. 3 (Autumn), 2005, pp. 512-548.
[1680] GRIFFITHS, A., "The practical challenges of implementing algorithmic regulation for public services", in YEUNG, K. & LODGE, M., (eds.), *Algorithmic Regulation*, Oxford, Oxford University Press, 2019, p.152.
[1681] GRIFFITHS, A., "The practical challenges of implementing algorithmic regulation for public services", *cit.*, 2019, p. 152.
[1682] CASTRO, D. & MCLAUGHLIN, M., "Ten ways the precautionary principle undermines progress in artificial intelligence", *Information Technology & Innovation Foundation*, 4th February 2019. Available on 14th May 2020 at: https://itif.org/
[1683] EU COMMISSION, "White Paper on Artificial Intelligence…", *cit.*, 2020.
[1684] DATENETHIKKOMMISSION, "Gutachten der Datenethikkommission", 2019; DATENETHIKKOMMISSION, "Opinion of the Data Ethics Commission: Executive Summary", 2019.

reports and scholarly articles.[1685] Moreover, the Canadian Government has also passed the Directive on Automated Decision-Making which establishes a four level system of requirements depending on the risks presented by algorithms but which only applies to public automated decision-making systems.[1686] Said versions of risk-based approaches to algorithmic regulation take into consideration the different risks generated by algorithms, which were generally pointed out in the first chapter to the second part of the dissertation and include risks to personal autonomy, biases in the general sense and harms to privacy rights amongst other possible damages.[1687]

The risk-based approach put forward in the following pages mainly focuses on the risks that algorithms generate to the rights to equality and non-discrimination and equality as a core value of democratic societies but can also be used as a blueprint to address the regulation and control of other risks of algorithms. This is important because although the proposal might seem to advocate for an excessively pro-regulatory framework, it does so because it only addresses algorithms that have shown to cause harms to the fundamental rights to equality and non-discrimination which, given their superior value, must be prioritised over the possible and, in many cases, unproven losses in efficiency that might be derived from placing certain administrative burdens on algorithm developers and controllers. This does not mean that other fundamental rights, such as the freedom to conduct a business, must not be taken into consideration, but that the risks and harms that algorithms have proven to generate to the principle of equality and to the rights to equality and non-discrimination require setting some limits to other rights.

There are many algorithmic applications that should not be subjected to the stricter oversight and accountability mechanisms that are suggested for algorithms that have been proven or show a clear risk for the rights to equality and non-discrimination. This reduction in administrative burdens is precisely what is sought through the implementation of a risk-based regulatory approach. However, the algorithms analysed in this dissertation risk harming individuals' right to equality and non-discrimination and perpetuating the structural

[1685] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019; SAURWEIN, F. *et al.*, "Governance of algorithms: options and limitations", *Digital Policy, Regulation and Governance*, vol. 17, No. 6, 2015, pp. 35-49.
[1686] Canadian Directive on Automated Decision-Making, 1st April 2019.
[1687] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, p. 39; SAURWEIN, F. *et al.*, "Governance of algorithms…", *cit.*, 2015, pp. 37-41.

discrimination and negative stereotyping of traditionally oppressed and vulnerable social groups and should therefore be subjected to stricter requirements.

Drawing from this idea and particularly inspired by the risk approach presented by the German Data Ethics Commission[1688] a three-tier system for algorithms that have proven harmful for the rights to equality and non-discrimination is presented in the following pages. In addition to this three-tier system, two more levels should be established corresponding to non-risky algorithms and automated systems that generate an unacceptable level of risks and must therefore be banned for the time being.

4.2.1. The three (plus two) tier system

4.2.1.1. Prohibited algorithmic systems

Firstly, there are a series of products or instruments that should be outright prohibited due to the very significant risks they cause on the fundamental rights of humans. An example of this type of prohibition, although structured quite lightly, is the one contained in article 22 of the GDPR. It is however important to recall that, while this prohibition is absolutely necessary as a minimum safeguard, it is nothing more than that, for the exceptions to the prohibition of solely automated decision-making allowed by the article allow for a wide array of fully automated decision-making (including profiling) systems to be legally implemented.

However, there are specific categories of automated systems that should be outlawed. In this sense, systems should be classified by their level of complexity and in relation to the purposes that they are deployed for. For instance, algorithms used towards particularly risky aims that entail significant interferences in the fundamental rights of individuals should not be opaque.

Additionally, there are certain purposes for which the use of automated systems should be banned. It is the case of using profiling systems to target advertisements of products and services that entail borderline or directly abusive conditions, that is, predatory advertising. Certain types of products and sometimes even whole firms, particularly in the financial services sector, should be flagged and tested and the automated systems used in advertising

[1688] DATENETHIKKOMMISSION, "Gutachten der Datenethikkommission", 2019, pp. 177-180.

them should be tested in order to ensure that the main recipients of said ads are not vulnerable populations.

Finally, there are also certain prohibitions that should be placed on public authorities outsourcing algorithms. There are certain activities that have an inherently public nature and that cannot be outsourced to private firms. This includes general law enforcement, including systems aimed towards detecting any form of illicit activity and establishing any form of sanctions.[1689] The fact that inherently public tasks are sometimes developed with the cooperation of private actors (for example, private expert opinions can be requested in the development of legal instruments), cannot, in any way, serve as the basis for authorising private parties to make the final decisions on inherently public tasks.

4.2.1.2. High-risk algorithmic systems

This category of algorithmic systems should include algorithms used in employment contexts for recruitment, promotion and any other purposes that influence pay. All systems that in any way influence individuals' access to credit and directly or indirectly set insurance primes would also be comprised under this category as well as algorithms used in medicine.

Algorithms used by public bodies to influence decisions made within the framework of public coercive activity, from administrative sanctions to sentencing, should also be subjected to the requirements set in this section. Given the role that public service provision and in particular, welfare services and public aid, has in promoting equality and, particularly, in redressing historical inequalities, the automated systems used for said purposes should also fall under the high-risk category, for any mistake or bias in the system will lead to harms to equality both as an individual right and superior value of democratic societies. The risks generated by public algorithms on vulnerable populations and members of disadvantaged groups, are especially enhanced if we consider that many of these systems are increasingly used as surveillance tools to detect fraud in social security and other forms of public assistance. When the aforementioned public tasks are outsourced to private parties, the algorithms used should also qualify as high-risk systems.

[1689] RANCHORDÁS, S. & SCHUURMANS, Y., "Outsourcing the welfare state…", *cit.*, 2020, pp. 31-34.

Finally, as a result of the "with great power comes great responsibility" rule that is applied to public administrations, all automated systems used by large tech companies[1690] that affect individuals by, for instance, limiting their choices or influencing their views, should be subjected to the highest level of requirements that are explained in the following pages. First and foremost, the enormous economic capacity (and uncompetitive behaviour)[1691] that characterises these firms, voids the innovation-hampering argument wielded by certain sectors of academia. Second, big tech effectively exercises regulatory powers over the digital environment by setting up its architecture and shaping and controlling the behaviour of cyberspace users.[1692] A very significant number of interactions and individual activities currently take place in the digital environment, providing these firms with an unprecedented level of power attributed to private sector actors over individuals' and society's opinions, attitudes and views.[1693]

Furthermore, large technological companies do not only act as regulators in online environments but are also increasingly exercising regulatory powers in the real world. We are witnessing a surge in the number of public-private partnerships set up in order to provide public services in smart cities.[1694] Although outsourced services should, in any case, abide by the same transparency requirements as automated systems developed by public administrations, it is also highly relevant to acknowledge the great degree of power that the participation of big tech in public activities provides these companies with. The harms these companies cause to the fundamental rights of individuals, including individual's freedom and autonomy, and basic values of democratic societies and the lack of effectiveness that antitrust regulations have shown thus far in limiting their powers, calls for greater intervention in these firms in order to deal with the dangers they pose.

i) *Administrative testing, documentation and general explanation requirements*

Systems in the top tier would require premarket administrative approval[1695] to operate based on an Impact Assessment carried out by each the public authority in charge of supervising

[1690] Apple, Alphabet, Amazon, Facebook, Match Group, Microsoft, Mozilla Foundation, etc.
[1691] KIM, N. S. & TELMAN, D. A., "Internet giants as quasi-governmental actors and the limits of contractual consent", *cit.*, 2015, pp. 723-770.
[1692] LESSIG, L., *Code: version 2.0*, *cit.*, 2006, p. 136.
[1693] KIM, N. S. & TELMAN, D. A., "Internet giants as quasi-governmental actors and the limits of contractual consent", *cit.*, 2015, p. 765; NOBLE, S. U., *Algorithms of Oppression…*, *Op. cit.*, 2018.
[1694] RANCHORDAS, S., "Nudging citizens through technology in smart cities", *cit.*, 2019, p. 14.
[1695] DATENETHIKKOMMISSION, "Gutachten der Datenethikkommission", 2019, p. 179.

algorithms. The scope of action of the European and national public authorities should be determined mainly following the remit of existing data protection authorities.[1696] Impact Assessments should analyse training and test data and the way in which the algorithm works with both datasets. In addition, given the constant development of the technologies used in automated systems, frequent audits would also have to be carried out by the relevant public body. Audits should compare the results for training and test data against real world data as well as any elements (instructions) contained in the algorithm that might change overtime as a result of self-learning and modify its behaviour.

The initial Impact Assessment and subsequent audits should establish whether the decision made by the algorithm is discriminatory and, in that case, whether it is directly or indirectly based on membership to especially protected groups. The Impact Assessment and audit results should be made fully available to the public on the website of the firm or institution operating the algorithm and should also be linked through the algorithmic regulatory body's website. A general explanation of the algorithm should also be made public that should include the type of automated system used, the sources from which the data is retrieved, a general description of the measured features, the class labels and target variable and their definitions and the general logic underlying the system.

In addition, while full transparency shall not necessarily be required for algorithms used by private firms, that is, source codes will not have to be published, processors and controllers will have to privately document every detail of the algorithm, including its source code and the weight and value attributed to each measured variable and provide said documents both for the initial Impact Assessment and subsequent audits as well as to any authorised certification organisms and other organisations involved in consumer protection and the defence of human rights that request access to them.

This system should be based on the existing DPIA and audit system contained in the GDPR. In fact, this proposal does not establish more requirements than the ones that are already theoretically set in article 35 of the GDPR, it does however aim to ensure the effectiveness of an accountability system that is currently failing and, obviously, to extend protection beyond the right to privacy.

[1696] Establishing the exact remit of each Authority will bring about complex issues given that many of these systems are created in a country but are sold and operate in many jurisdictions. This discussion falls outside the scope of this dissertation.

ii) *Justification and explainability requirements*

The way in which the algorithms included in this category work should also have to be properly justified. As it was already indicated in Part II, one of the GDPR's most significant shortcomings is the fact that it focuses on protecting input personal data but does not offer sufficient protection regarding the inferences that may be drawn when said data is processed[1697] and especially the inferences that result from the processing of certain types of data that are not covered by the GDPR, such as anonymised data, which can have very significant impacts on the lives of individuals.[1698]

By recognising the "right to reasonable inferences"[1699] in certain sectors in which automated, semi-automated decision-making and/or profiling are used, algorithm designers will have to ensure that the results yielded by the system are justified.[1700] Part of the scholarship has defended that, for the way in which an algorithm works to be justified, it must satisfy intuitiveness requirements, that is, the relationships the algorithm establishes between different sets of variables must be understandable and the relevance of each measured feature with regard to the decision must also be reasonable.[1701] Moreover, this was the approach suggested by the US Federal Reserve as it established in its 2011 Guidance on model risk management that "outcomes analysis… can also include expert judgment to check the intuition behind the outcomes and confirm that the results make sense".[1702]

Some commentators have argued that subjecting algorithms to such a degree of transparency and explainability might reduce their accuracy and stifle innovation.[1703] They contend that some of the automated systems that are being currently used have reached such a degree of complexity that explaining them and, more importantly, ensuring that the explanation provided is intuitive, is not possible and will therefore require using simpler algorithms that will reduce the level of accuracy which could, in some cases, lead to an increase in biased

[1697] WACHTER, S. & MITTELSTADT, B. D., "A right to reasonable inferences…", *cit.*, 2019, p. 572.
[1698] SCHREURS, W. *et al.*, "Cogitas, ergo sum…", *cit.*, 2008, pp. 241-257; WACHTER, S. & MITTELSTADT, B. D., "A right to reasonable inferences…", *cit.*, 2019, p. 576.
[1699] WACHTER, S. & MITTELSTADT, B. D., "A right to reasonable inferences…", *cit.*, 2019, pp. 494-620.
[1700] *Idem*, p. 580.
[1701] GRIMMELMANN, J. y WESTREICH, D., "Incomprehensible discrimination", *cit.*, 2017, pp. 175-176; KIM, P. T., "Data-driven discrimination at work", *cit.*, 2017, pp. 922-923; WACHTER, S. & MITTELSTADT, B. D., "A right to reasonable inferences…", *cit.*, 2019, p. 581.
[1702] BOARD OF GOVERNORS OF THE FEDERAL RESERVE SYSTEM, OFFICE OF THE COMPTROLLER OF THE CURRENCY, "Supervisory Guidance on Model Risk Management", SR Letter 11-7, 2011, pp. 13-14.
[1703] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, pp. 1085-1139; CASTRO, D. & MCLAUGHLIN, M., "Ten ways the precautionary principle undermines progress in artificial intelligence", *cit.*, 2019.

and discriminatory outcomes.[1704] In addition, if firms can only use simple algorithms they will have no incentive to improve automated systems. Considering this statement can be true in some cases, it is necessary to come up with a solution in order for said algorithms to offer sufficient justifications.

A similar system to BAROCAS and SELBST'S proposal for "documentation as explanation" should be made mandatory for all algorithms. This system would be specially designed to ensure that even those algorithms that cannot be explained are subjected to an adequate *ex ante* control.[1705] Regardless of explainability, source codes will have to be offered in all cases. That is, they will not have to be published, but provided to the Authority. Moreover, when explanations cannot be delivered, the documentation system must ensure that a more explainable algorithm could not be used and that the closest thing to an explanation is provided, for example, through counterfactuals.[1706]

The documentation system should focus on answering three sets of questions regarding the way in which the algorithm works. The first set should determine the aim of the system, ensuring it is working the way it is supposed to, that is, it satisfies its functional specifications, and that a more explainable model cannot be used in order to achieve the same aim with a similar degree of accuracy. Once these elements have been considered, a query into the choices made by algorithm developers should be carried out, forcing them to justify the way in which data was collected and examples were labelled and the choices made regarding target variables, relevant features, how outliers were treated how the data was partitioned for testing and the degree to which the algorithm was tuned.[1707] The final set of questions should establish whether the results provided by the algorithm are automatically adopted or if there is any kind of human intervention.[1708] In addition, the organisation that developed the algorithm should also be examined in order to determine its composition and if any specific training on ethics is carried out. The organisation deploying the algorithm should be examined in those cases in which the system was used for semi-automated decision-

[1704] HUME, K., "When is it important for an algorithm to explain itself?", *Harvard Business Review*, 6th July 2018. Available on 16th May 2020 at: https://hbr.org/.
[1705] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, pp. 1129-1138.
[1706] KLEINBERG, J. *et al.*, "Discrimination in the age of algorithms", *cit.*, 2018, p. 145; WACHTER, S., MITTELSTADT, B. D. & RUSSELL, C., "Counterfactual explanations without opening the black box…", *cit.*, 2018, pp. 841-887.
[1707] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, p. 1131; LEHR, D. & OHM, P., "Playing with the data…", *cit.*, 2017, pp. 683-700.
[1708] BAROCAS, S. & SELBST, A. D., "The intuitive appeal of explainable machines", *cit.*, 2018, p. 1132.

making in order to assess the level of discretionality that employees have in making the final decision and the criteria and training received in order to do so.[1709]

The algorithm should obviously also be tested for discriminatory results by providing it with real-world data of members and non-members of different vulnerable groups, ensuring that an intersectional perspective is adopted when said tests are being carried out. If the algorithm yields discriminatory results it will be necessary to establish whether said differential treatment is lawful or does not respect the minimum levels of equality set for that particular algorithm. In some cases it will be necessary to require intuitive explanations and for the algorithm's reasoning to be evaluated both from a quantitative and qualitative perspective. For example, an algorithm used to determine individuals' creditworthiness will discriminate based on socioeconomic status but it will be necessary to determine whether the difference in creditworthiness between individuals in different socioeconomic positions is justified or whether it is disproportionate.

However, it is vital to ensure that this alternative system does not provide controllers and processors with a way out of offering intuitive explanations when possible. Intuitive explanations are necessary because they provide automated systems with a higher degree of perceived legitimacy and prevent dehumanising the individuals being subjected to their decisions by speaking a language they can understand. More importantly, intuitive explanations enhance the effectiveness of the rights to be heard and contest decisions.

iii) *The proportionality analysis of pre-market authorisations*

The risk-based proposal serves a legitimate aim in that its objective is to protect fundamental rights and public interests such as the right to equality and non-discrimination, the right to data protection, equality as a general principle that grounds the actions of public institutions and the freedom and autonomy of individuals.

Pre-market authorisations are adequate to the aims pursued in that they ensure that the rules and mandates that certain processes that affect individuals must abide by are applied and introduced into the algorithm carrying out said processes. If employers cannot discriminate on the grounds of sex a pre-market authorisation will ensure that the algorithm used by

[1709] *Ibidem.*

human resources in a certain company does not directly or indirectly discriminate on the grounds of sex.

Pre-market authorisations are necessary for algorithms that generate the highest risks for fundamental rights because *ex post* controls would generate a time-frame in which it would be possible to harm individuals' rights to equality and non-discrimination, as well as other rights or values. In addition, the effectiveness of *ex post* public controls must be put into question since, in many cases, public administrations end up lagging behind and not effectively exercising supervisory powers. While this could be solved through political willingness to implement and carry out effective inspections on regulated products and activities, given the harms that data processing technologies have proven to cause to fundamental rights and public interests, it does not seem advisable to leave the door open to the perpetuation of the current system in which algorithms are used for purposes that significantly impact individuals and their rights without hardly any form of oversight.

Finally, attending the *strictu sensu* proportionality test, if pre-market approval processes are efficiently structured, they should not entail significant harms for innovation or for the freedom to conduct a business. Obviously, in the case of algorithms developed by big tech (but also other large companies), their capacity of assuming greater bureaucratic requirements renders any argument regarding how these pre-market authorisations could stifle innovation, void. However, in any case, when the risks generated to the fundamental rights and freedoms of individuals are high, protection to said rights through a system of market pre-authorisations outweighs other competing interests.

Competing rights and interests, including intellectual property rights and the freedom to conduct a business, must obviously be taken into consideration and the system of prior authorisations must be carefully crafted in order to not unnecessarily and disproportionately overburden medium and small firms. However, considering the institutional system built by the European Union is mainly aimed towards the protection of economic freedoms, pre-emptively protecting other rights and interests, namely the rights to equality and non-discrimination, when the risks to said rights have been recognised to be very high, is necessary in order to rebalance a system that, while recognising the importance and mandates set by the rights to equality and non-discrimination, mainly prioritises economic freedom and efficiency. Establishing a few limits on the freedom to conduct a business does not mean that the principle of equality generally outweighs said freedom, but that in certain particular

contexts it is necessary to establish certain minimum corrections in favour of equality in order to ensure the rights and freedoms of the members of disadvantaged groups. This is further justified if we consider the generally asymmetric, informational and otherwise, positions of the parts in automated decision-making processes.

It is finally relevant to point out that the GDPR establishes mandatory DPIAs prior to deployment when systems generate high risks for the rights and freedoms of natural persons. Hence, the need to carry out a risk assessment prior to deployment has already been recognised by European regulators. However, as it has already been stated several times, the oversight system of collaborative governance included in the GDPR has not proven to be effective.

iv) *Specific requirements for public sector algorithms included in this category*

The algorithms employed by public administrations, whether publicly or privately developed, should be subjected to full transparency and publicity requirements. Unlike in the case of algorithms deployed by the private sector, the full transparency of which will only be available to certain actors, the algorithms used for public purposes included in this category shall be made fully transparent and available to de public in the relevant body's website.

In this sense, it is relevant to point out that the algorithm register established in Amsterdam could serve as an example on which to base other websites on which to publish information on algorithms used by the public sector. The city of Amsterdam has recently established an algorithm register in which it is possible to find information on the automated systems used by the city's public sector. They include information on different aspects of the algorithm, including whether the system is at risk of discriminating. For instance, with regard to the algorithm used for detecting housing rental fraud risk, which is still in its pilot version, the website indicates the risks that the system might generate outputs that discriminate against certain groups and refers to the AI fairness toolkit that is being used in order to detect and avoid discrimination during the pilot.[1710]

In addition, as it was already argued, these systems effectively operate as regulatory instruments,[1711] meaning citizens should have the same guarantees with regard to algorithmic

[1710] ALGORITHM REGISTER, "Housing rental fraud risk", 2020. Available on 9th October 2020 at: https://algoritmeregister.amsterdam.nl/
[1711] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, pp. 223-270.

systems as they do with other regulatory instruments, including algorithmic (normative) planning, pre-establishing a set of values the system must follow, public participation in the algorithm's development and *ex ante* and *ex post* reviews of the rule.[1712] In addition, with regard to the *ex post* control of publicly used algorithmic systems, in all instances in which algorithms may have in some way influenced the decision, individuals' right to review and contest the decision should include the possibility of fully reviewing the algorithm.

The only instances in which it is possible to justify that full public transparency is not provided is in cases in which the risk of gaming is real. The typical examples cited are algorithms used to red-flag individuals, whether it is within the framework of taxation plans in which automated systems signals citizen's that are suspected of tax fraud or in airport security checks.[1713] Hence, the features that cause the algorithms employed by taxation agencies to signal specific citizens as suspect of committing tax fraud are not currently made public.[1714] In order to ensure that these algorithms are not discriminatory they should however be made fully available to the public supervisory body and, whilst not making the source code public, following the system already established for tax inspection plans, general public explanations should be published.[1715]

Algorithms used for public coercive purposes should also be subjected to stricter explainability and justification requirements. An indirect way in which this can be done is through the obligation that the public decisions influenced by automated systems are not exclusively based on the recommendations provided by the algorithmic system, which would basically mean implementing the prohibition set in article 22 of the GDPR. That is, instead of requiring the algorithm to be fully justified, said requirement is set for the final decision, which must be sufficiently supported not just by the results yielded by the algorithm but also by other forms of proof that, alongside the algorithms recommendation help build a strong case for suspecting an individual to be carrying out illicit activities.[1716]

[1712] *Idem*, pp. 252-253.
[1713] BAYAMLIOĞLU, E., "Transparency of automated decisions in the GDPR…", *cit.*, 2018, p. 18; ZARSKY, T., "Transparent predictions", *cit.*, 2013, pp. 1553-1554.
[1714] CALDERÓN CARRERO, J. M., "El encuadramiento legal y límites del uso de herramientas de inteligencia artificial con fines de control fiscal: Análisis de la decisión del Consejo Constitucional francés de 27 de diciembre de 2019 (Décision n.º 2019-796 DC), sobre la Ley de Presupuestos 2020", *CEF Revista de Contabilidad y Tributación CEFLegal*, Nº. 444, 2020, p. 125.
[1715] BOIX PALOP, A., "Los algoritmos son reglamentos…", *cit.*, 2020, p. 265.
[1716] CALDERÓN CARRERO, J. M., "El encuadramiento legal y límites del uso de herramientas de inteligencia artificial con fines de control fiscal…", *cit.*, 2020, p. 121.

While some regulators seem to consider that subjecting the automated decision to human review is sufficient to ensure that right to due process is respected and the decision is fair, it is vital to acknowledge, as it has been repeatedly pointed out throughout this dissertation, the great deal of influence that algorithmic system recommendations have on human decision-making. It is therefore necessary for the automated decision to also be sufficiently justified, especially if it affects the freedom of individuals, as it happens in the exercise of all types of public coercive powers. The weight the algorithmic decision should have on the final decision should be pre-established, taking into consideration, amongst other issues, the risk of automation bias.[1717] In those cases in which the algorithm plays a very important role in the final decision it should not just be based on correlations but should also be tested for causality. In addition, given that all algorithms in this category are to be subjected to regular audits by the supervisory authority, the real influence that the automated system has on the decision will be appreciated through said audits, which could be used as the basis to require greater degrees of explanation.

4.2.1.3. Medium-risk algorithmic systems

This category would, for example include targeted advertising and personalised pricing algorithms. These systems should undergo a prior certification process carried out by public bodies or private firms authorised by the corresponding Authority as certification entities. Simpler requirements than the ones set for high-risk algorithmic systems should be set. For example, in the case of targeted advertising, the class labels and target groups would have to be revealed in order to ensure that the algorithm is not used for predatory advertising purposes. The algorithm should also be run in order to determine that no discriminatory or stereotyping results were being produced. However, other elements should not have to be initially tested in order for the algorithm to be approved and could be simply subjected to *ex post* audits.

System source codes and all other relevant documentation should be provided to the supervisory authority during the authorisation procedure and updated every few months. The supervisory authority should carry out an initial *ex post* audit in order to fully analyse all the different elements that may risk making the algorithm discriminatory. After the first audit, regular reviews should also take place but could be substituted by certification systems

[1717] CITRON, D. K., "Technological due process", *cit.*, 2008, pp. 1271-1272.

carried out by authorised certification bodies, however, even if this option is taken up, every few years the system should be audited by the supervisory authority to ensure that it still complies with all requirements and the certification mechanisms are working properly.

4.2.1.4. Low-risk algorithmic systems

This category should include algorithms used in GPS devices or for dynamic price setting, as long as the latter do not influence pay, as they do in the case of algorithms used in sharing economy platforms. However, considering there are still risks that dynamic pricing discriminates against the populations of vulnerable areas, it is also important that these systems are subjected to some form of control. Authorised certification bodies should be in charge of testing and certifying the validity and legal compliance of these systems.[1718]

Taking up the basis set by article 42 of the GDPR with regard to certification mechanisms, new regulatory instruments that address automated systems in a more comprehensive manner should establish certification as mandatory. However, considering the lower risk of these systems, control could be carried out once they have been deployed. Certificates should be renewed every few years.

4.2.1.5. Non-risky algorithmic systems

Finally, non-risky algorithmic systems, such as the ones used in all types of vending machines should simply undergo the normal quality control requirements that are set for these types of products.[1719]

4.2.1.6. System enforcement

One of the problems with controlling algorithms is that their use is highly disperse and difficult to identify. In order to ensure the effectiveness of the risk-based system, it should be implemented in a progressive manner. The first step would be to only subject organisations that are forced to designate or have designated a data protection officer (article 37 of the GDPR) to this system. In addition, the Authorities charged with algorithmic supervisory powers should designate teams tasked with detecting particularly risky and harmful uses of

[1718] DATENETHIKKOMMISSION, "Gutachten der Datenethikkommission", 2019, pp. 178-179.
[1719] *Idem*, p. 178.

algorithms and processors and controllers that are not under the radar of public authorities. This initial step could help to kick-start the system the proposed risk-based system that would, in a progressive manner, move on to control algorithms created and used by organisations that are not required or have not designated a data protection officer.

4.3. PUBLIC PROCUREMENT AS A MECHANISM TO PREVENT THE RISKS OF THE PUBLIC AND PRIVATE USE OF ALGORITHMS

Several proposals and guidelines have been put forward in order to set specific requirements that the private organisations contracted by public institutions must meet when automated systems are acquired.[1720] In this context, it is important to highlight that public procurement has been used as a mechanism to advance social policy objectives for quite some time.[1721] Hence, setting specific mandates that the private firms that procure software systems to public administrations must comply with can help to indirectly promote said firms to adopt certain internal organisational policies that help to create better and less discriminatory algorithms. For example, private parties that enter into contracts with public administrations for the procurement of software should in all cases comply with diversity requirements. In addition, by forcing the algorithms developed by private firms to pass equality impact assessments or to be accountable, said companies will be compelled to generally develop all systems in said way whether they will be deployed by public or private actors.

4.4. ESTABLISHING A "BEST AVAILABLE TECHNIQUES" REGIME

The "best available techniques" mandate was already suggested to be applied to cases of algorithmic discrimination when determining whether the use of a discriminatory algorithm can be deemed necessary. For the "best available techniques" mandate to be applicable, it is necessary to establish a European reference system that member states and courts can refer to when evaluating the extent to which, taking into consideration technological advances and the affordability of fair algorithms, which algorithms should be used for which purposes and the degree of transparency, equality and respect for fundamental rights and public interests than can be achieved by different firms in different sectors.

This system should be inspired by the "best available techniques" reference documents and exchange of information set out by article 13 of the 2010 Directive on industrial emissions,

[1720] KOENE, A. *et al.*, "A governance framework for algorithmic accountability and transparency", *cit.*, 2019, pp. 62-63; UK GOVERNMENT, "Guidelines for AI procurement", 8th June 2020.; WORLD ECONOMIC FORUM, "Guidelines for AI procurement", 2019. Available on 30th May 2020 at: https://www.weforum.org/

[1721] MESTRE DELGADO, J. F., "El tratamiento jurídico de la discapacidad en la ley de contratos del sector público", *Anales de Derecho y Discapacidad*, No. 3, junio 2018, p. 88.

which works fairly well.[1722] Enhancing communication between member states, institutions and sectors developing and using algorithms for different purposes will help to build an overarching system of algorithmic oversight and control.

4.5. USING ALGORITHMS TO DETECT DISCRIMINATION

As it has already been stated on several occasions, there is a growing body of literature focused on the creation of fair algorithms. Algorithmic systems should be used in order to detect instances of discrimination and areas of society in which public institutions should intervene in order to redress persisting inequalities. If used mindfully, algorithms could, in fact, become an asset in the pursuit of a more equal and just society. For example, algorithms have been used in order to detect hate speech against immigrants and women on Twitter.[1723] In this sense, algorithms could also be used in order to detect instances of systemic and institutional discrimination contained in apparently neutral social structures. Moreover, if properly articulated, algorithms could also be used in order to detect and remove instances of discrimination in other automated systems,[1724] therefore rendering control and oversight procedures much more efficient.

4.6. EMPOWERING INDIVIDUALS THROUGH UNDERSTANDABLE INFORMATION: CHOICE ARCHITECTURES

One of the problems of the current framework of protection against the risks generated by data processing technologies, which largely burdens individuals with protecting their own rights and interests, is that individuals are not really aware of the data they are sharing and the risks that sharing their data can entail. While an effective oversight and accountability system is necessary, it is also essential to empower individuals and promote general literacy regarding the issues concerning the data services sector.

In this sense, an essential action that must be taken is returning individuals their freedom in deciding how to protect their privacy. Freedom only exists if a person has the possibility of making choices. The way in which the data protection system is set up allows firms that

[1722] REVUELTA PÉREZ, I., "Mejores técnicas disponibles": un singular sistema de regulación ambiental", *Revista Catalana de Dret Ambiental*, vol. 10, No. 1, 2019, pp. 1 -34.
[1723] GARIBÓ I ORTS, O., "Multilingual detection of hate speech against immigrants and women in twitter at semeval-2019 task 5: frequency analysis interpolation for hate in speech detection", *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 460-463.
[1724] FELDMAN, M. *et al.,* "Certifying and removing disparate impact", *cit.*, 2015, pp. 259-268.

provide digital services to create choice architectures that manipulate them into sharing their data whether it is as a result of being overwhelmed by too much information, by being repeatedly asked every time they click on a tab whether they want to remove their consent, by directly forcing them to share their data in order to access the service, or even by making data-sharing option buttons more colourful or bigger. Public and private organisations should be forced to set up very simple and visible messages which provide individuals with the possibility of deciding one single time whether they want to share their data or not.

4.7. INCREASED COMMUNICATION BETWEEN DISCIPLINES AND ESTABLISHING GENERAL PRINCIPLES UPON WHICH TO CONSTRUCT AUTOMATED SYSTEMS

The first section in this chapter addresses the frictions that exist between traditional and algorithmic forms of regulation and the need to make very concrete policy and regulatory choices that, in traditional regulatory settings, are specified on a case-by-case basis.

Algorithms work in ways and present challenges that require a completely different mindset and framework of action to the one legal operators and scholars are used to. Conversely, the scientific method employed by programmers and IT experts fails to grasp the need to develop algorithms for *ex post* oversight whenever possible and, more importantly, some of the core principles and values that democratic societies are founded on and the need to protect the rights of individuals, and in particular, of ensuring the freedom of vulnerable and historically disadvantaged groups. Increased collaboration between disciplines must therefore be fostered to ensure that algorithms are constructed as tools that improve our lives and not as surveillance and control systems that treat the many as tools in service of the few.

It is not just important to foster collaboration but to consider introducing IT experts in areas in which algorithmic systems are being and will be increasingly used. In this sense, DE LA SIERRA, points to the possibility of including tech experts in courts and the need to train other intervening agents in the process on the way in which these systems work to ensure that cases in which the use of algorithms is contested or algorithms have intervened in some way, are properly judged.[1725] Moreover, tech experts must also receive specific ethics training in order to ensure that they are aware of algorithms' potential for damage and the limits and principles that these systems should be subjected to.

[1725] DE LA SIERRA MORÓN, S., "Inteligencia artificial y justicia administrativa: una aproximación desde la teoría del control de la Administración pública", *Revista General de Derecho Administrativo*, No. 53, 2020, pp. 11-12.

RESULTADOS Y CONCLUSIONES FINALES

1. Las cuestiones jurídicas derivadas de la creciente utilización de los algoritmos constituyen un problema cuyo ámbito de estudio y resolución ha de ser, necesariamente, supranacional

Las razones que hacen que el tema objeto de análisis de esta tesis doctoral se estudie desde una perspectiva supranacional son, por una parte, que los problemas y efectos de los algoritmos empleados en sistemas de decisión automatizada se desarrollan de manera globalizada y son muy similares o idénticos allá donde se producen y, por otra, que las principales respuestas normativas para su control son, y están llamadas a ser, supranacionales.

En primer lugar, los sistemas automatizados de toma de decisiones desarrollan sus funciones principalmente en el espacio digital, un espacio en el que las distancias físicas pierden su relevancia y que se caracteriza por proporcionar la interconexión de una parte muy importante de las y los habitantes del planeta. Estos sistemas se desarrollan en el marco de un sector económico, el de los servicios informáticos de la sociedad de la información, cuya actividad se presta preponderantemente de manera internacional. Es por ello que las partes intervinientes en cada proceso algorítmico de toma de decisiones se suelen hallar distribuidas a lo largo de diferentes Estados. Asimismo, las clases de efectos negativos que estos sistemas pueden tener para las personas físicas son más próximas entre sí por rama de actividad y sector en los que se emplean los algoritmos que por ubicación de la persona destinataria de la decisión automatizada.

En segundo lugar, las principales respuestas normativas que se han dado hasta la fecha, esto es, la regulación a través de la protección de datos, se han producido y desarrollado a nivel europeo. Es más, las recientes discusiones y propuestas normativas relacionadas con el desarrollo de un marco jurídico más completo para el control de los algoritmos se están produciendo, principalmente, también a esta misma escala. Así, si bien el legislador nacional de cada Estado sí deberá ajustar el uso de los sistemas automatizados a su propio ordenamiento jurídico interno en ámbitos concretos, la regulación y límites generales al uso de nuevas tecnologías de procesamiento de datos y de toma de decisiones automatizadas vendrá, en un futuro no muy lejano, y en la misma línea que el actual marco jurídico en materia de protección de datos, inevitablemente determinada por las instituciones europeas.

2. La discriminación algorítmica constituye un problema de carácter bicéfalo que incorpora i) los efectos de este fenómeno sobre los derechos a la igualdad y a la no discriminación y, ii) los problemas específicos derivados de la vulneración a los derechos fundamentales y a otros valores superiores propios de los Estados de Derecho democráticos, vehiculada a través de herramientas informáticas

El marco jurídico en materia de igualdad y no discriminación se encuentra llamado a proteger frente a cualquier clase de daño causado a los derechos fundamentales de los que se ocupa. Es por ello que resulta irrelevante que sean seres humanos o programas de software quienes tomen decisiones discriminatorias pues, mientras se produzca dicho efecto, deberán aplicarse las normas jurídicas que protegen los derechos a la igualdad y a la no discriminación. Asimismo, a nivel teórico, la discriminación algorítmica de las personas pertenecientes a grupos o colectivos que han sido históricamente relegados a un segundo plano en la construcción de las instituciones de poder también debe analizarse y entenderse en el marco de la continuidad de la discriminación estructural y específica sufrida por estos grupos y sus integrantes.

Ahora bien, es necesario atender al hecho de que la discriminación resultante de los procesos algorítmicos, y también la perpetuación de la desigualdad que puede resultar de la automatización de los sistemas de servicios y ayudas públicas, se producen a través de un fenómeno muy específico: el creciente desarrollo e introducción de tecnologías de procesamiento de datos y de automatización de decisiones. Es por ello que la discriminación algorítmica debe ser abordada también como uno más de los riesgos y daños generados por esta clase de sistemas y en el contexto de las características específicas de las decisiones automatizadas. En este sentido, es esencial no perder de vista la íntima relación entre los diferentes elementos y riesgos propios de los algoritmos. Por ejemplo, la falta de transparencia de estos sistemas es una de las cuestiones que más ha puesto la doctrina de relieve y que sin duda debe resolverse de cara al desarrollo de un sistema efectivo de detección y control de la discriminación algorítmica.

En definitiva, el análisis de las respuestas reguladoras actuales y futuras para la discriminación algorítmica debe realizarse tanto desde la perspectiva del marco jurídico en materia de igualdad y no discriminación como de los instrumentos de regulación dirigidos a combatir, de manera global, los diferentes problemas y riesgos surgidos como consecuencia

del creciente uso de algoritmos en los procesos de toma de decisiones que afectan a las personas.

3. La perpetuación de las estructuras sociales de discriminación continúa siendo una realidad que los ordenamientos jurídicos de los Estados occidentales reconocen formalmente y tratan de atajar mediante la introducción de herramientas de prevención y respuesta a las vulneraciones a los derechos a la igualdad y la no discriminación

Las sociedades occidentales se han construido, históricamente, sobre unas estructuras de subordinación de unos grupos en relación con otros. Así, con la aparición de las primeras democracias liberales, a las personas pertenecientes a los grupos subordinados les fue negada la condición de ciudadanía plena. En la actualidad y, en algunos casos, desde hace ya décadas, se reconoce la participación total de las personas pertenecientes a estos grupos o colectivos. Sin embargo, dicho reconocimiento y acceso formal a las estructuras e instituciones de poder no ha venido necesariamente acompañado de un disfrute en igualdad de condiciones de los privilegios y prerrogativas de los grupos que, históricamente, han ocupado el poder.

La igualdad constituye un valor superior y principio en el que las instituciones públicas de los Estados democráticos de Derecho deben basar su actuación. Asimismo, la igualdad y no discriminación se configuran como derechos fundamentales subjetivos de la ciudadanía de estos Estados, los cuales tienen la obligación negativa y positiva de proteger estos derechos subjetivos. Sin embargo, esta construcción parte de una aproximación eminentemente formal a la igualdad que no considera las estructuras subyacentes de discriminación que sitúan en una posición de desventaja a las personas pertenecientes a aquellos grupos que históricamente se han situado en una posición de subordinación.

En la actualidad los ordenamientos jurídicos de la UE y de sus Estados miembros reconocen el especial desvalor de las situaciones de discriminación por razón de una serie de categorías sospechosas. Asimismo, de manera creciente se ha ido ampliando el cuerpo de normas que reconocen estos derechos en las relaciones entre particulares y, en paralelo, una parte cada vez más importante de la doctrina defiende la eficacia horizontal inmediata o directa de los derechos a la igualdad y la no discriminación. Ahora bien, puesto que estas construcciones se basan fundamentalmente en perspectivas formales de la igualdad, el reconocimiento de una

especial protección frente a la discriminación por razón de sexo, raza, religión, etc., se basa, sobre todo, en la asunción del hecho de que estas son categorías (en principio) inmutables o que pertenecen al espacio inherente a la dignidad humana.

Así, la especial protección ofrecida en estas situaciones no se centra, o al menos no de manera principal, en el reconocimiento de la discriminación sistémica que dentro de cada una de estas categorías sospechosas coloca a las personas pertenecientes a un subgrupo (mujeres frente a hombres, minorías raciales frente a personas blancas, etc.) en una posición de desventaja social y de especial riesgo de trato desigual. Es por ello que, a pesar de que las perspectivas sustantivas de igualdad y no discriminación se van incorporando progresivamente a los ordenamientos jurídicos de los Estados democráticos y de Derecho, el ordenamiento jurídico únicamente reconoce y aborda la existencia de estas estructuras sociales de opresión de manera fragmentaria y, en demasiadas ocasiones, poco eficaz.

El reconocimiento y protección de los derechos a la igualdad y a la no discriminación encuentran importantes tensiones en relación con el libre desarrollo de la autonomía de la voluntad de los individuos y, en algunos casos, con la eficiencia entendida desde una perspectiva económica. Es más, en el sector privado, estos dos elementos (autonomía y eficiencia) tienden a conjugarse y cristalizar en la libertad de empresa, que puede verse limitada a través de la protección de la igualdad. Sin embargo, la libertad únicamente existe cuando las personas tienen verdadera capacidad de decidir, es decir, si tienen opciones. En cambio, si a una persona le es negado el acceso a un bien o servicio o no se le escoge para un puesto de trabajo porque pertenece a un grupo desaventajado, o por otros elementos que se encuentran vinculados a la pertenencia al grupo, se le está negando la libertad. Es por ello que la protección de la igualdad debe también ser entendida como protección de la libertad y, también, la razón por la cual, en ocasiones, una medida que aparentemente pueda suponer una limitación de la libertad, en realidad sirve para reequilibrar la balanza y garantizar la libertad de aquellas personas que, por su pertenencia a determinados grupos, parten con desventaja en algunos ámbitos de la vida social, económica y política.

Ahora bien, puesto que el ordenamiento jurídico europeo se asienta sobre unas estructuras cuyo objetivo fundamental ha sido garantizar la libertad económica, y puesto que dicha libertad económica ha sido históricamente construida desde la perspectiva de aquellos que ocupaban el poder, la protección de la igualdad, aunque reconocida, carece generalmente de la profundidad requerida para atajar, desde la raíz, la discriminación sistémica existente en la

sociedad. En el marco de la Unión Europea la libertad de empresa se protege por defecto y la protección de los derechos fundamentales a la igualdad y no discriminación se sitúa, en realidad, en un plano secundario. Ello se evidencia sobre todo con la evolución de la jurisprudencia del TJUE, en la que es manifiesta una creciente priorización de la libertad de empresa cuando esta se enfrenta con los derechos a la igualdad y la no discriminación.

4. La clasificación de los instrumentos jurídicos para la protección frente a la discriminación algorítmica debe basarse en la construcción de la doctrina estadounidense que diferencia entre instrumentos de anti-clasificación y anti-subordinación

El marco jurídico en materia de igualdad y no discriminación se compone de una serie de instrumentos de protección preventiva y reactiva frente a las situaciones de discriminación. Dentro de las herramientas de protección preventiva frente a la discriminación, la distinción que realiza la doctrina estadounidense entre instrumentos de anti-clasificación y anti-subordinación resulta útil de cara al análisis de la discriminación algorítmica, sobre todo con respecto a los paralelismos que pueden dibujarse entre la normativa de protección de datos y los instrumentos de anti-clasificación.

Los instrumentos jurídicos de anti-clasificación se refieren a aquellas normas que prohíben la consideración de las categorías especialmente sospechosas en los procesos de toma de decisión. Esta estrategia se encuentra íntimamente ligada al artículo 9 del RGPD, que prohíbe el procesamiento de categorías especiales de datos personales, las cuales incluyen el origen racial, las convicciones religiosas o las opiniones políticas.

Por su parte, los instrumentos jurídicos de anti-subordinación reconocen y pretenden revertir aquellas estructuras sociales que sitúan a las personas pertenecientes a determinados grupos o colectivos en una situación de desventaja. Entre los mecanismos incluidos en estos instrumentos se encuentran, por ejemplo, las acciones positivas y las iniciativas e instituciones que actúan a partir de la transversalización de la igualdad.

Con respecto a los instrumentos jurídicos reactivos frente a las situaciones de discriminación, ha de hacerse referencia a las prohibiciones de discriminación directa e indirecta. Las prohibiciones de discriminación operan también, en cierta medida, como mecanismos de prevención al articular mandatos que pretenden evitar la toma de decisiones discriminatorias.

Sin embargo, dado que los efectos de dichas prohibiciones tienen lugar como mecanismos *ex post* de identificación y resarcimiento de vulneraciones al derecho a la no discriminación, deben ser tratados y examinados como instrumentos de reacción frente a las situaciones de discriminación.

5. Los sistemas automatizados (y semiautomatizados) de toma de decisiones, que son empleados de manera creciente en todo tipo de áreas y sectores, reproducen las tradicionales estructuras de discriminación y desventaja que afectan a las personas pertenecientes a grupos históricamente oprimidos

La automatización de procesos no es una novedad pues, ya desde hace décadas, se emplean programas de *software* para la realización de actividades relativamente sencillas como pueda ser, por ejemplo, el cruce de bases de datos de diferentes Administraciones públicas. Sin embargo, lo que sí es novedoso y supone un cambio no sólo cuantitativo sino, a partir de cierto punto, también cualitativo, es la enorme y creciente capacidad computacional actualmente poseída por los sistemas de procesamiento de datos y automatización de procesos de toma de decisión, así como el aumento de su aptitud para el autoaprendizaje.

Tanto el sector público como el sector privado emplean algoritmos de manera creciente en todo tipo de procesos de toma de decisiones que, de manera directa o indirecta, afectan a los derechos y situaciones materiales de las personas. En el sector privado, podemos encontrar abundantes ejemplos de uso de sistemas automatizados en la selección de personas candidatas a puestos de trabajo, en la determinación de la capacidad crediticia de las personas físicas o en la determinación de los resultados de las búsquedas de información realizadas en plataformas digitales. En el sector público, los algoritmos se emplean en la configuración de los sistemas de ayudas públicas, en la provisión de servicios públicos y también en la actividad desarrollada por los cuerpos y fuerzas de seguridad, así como en otras clases de actividades de limitación, con una creciente incidencia en los procesos de análisis de datos y detección de pautas que permitan identificar conductas, o incluso preverlas, dentro de las actividades de inspección y control.

En este marco de creciente utilización de sistemas automatizados, se ha ido llamando la atención, en los últimos años, sobre la aparición de múltiples casos de discriminación algorítmica y de situaciones en las que la automatización de procesos contribuye a perpetuar las situaciones de vulnerabilidad de algunos grupos de la población. Así, se ha puesto de

relieve que los procesos de desarrollo y puesta en funcionamiento de los algoritmos ofrecen una amplia serie de posibilidades para la incorporación de sesgos que perjudican a los miembros de grupos desaventajados.

Una de las formas en las que se pueden crear algoritmos que reproduzcan las tradicionales estructuras de discriminación de miembros de grupos desaventajados es a través de la selección y peso otorgado a las variables empleadas por los algoritmos para la medición y predicción del objeto del que se ocupan. La forma en la que se da prioridad a unas variables sobre otras en la medición de los fenómenos que los algoritmos se encargan de predecir puede condicionar y sesgar los resultados de estos sistemas. Por ejemplo, si un sistema de calificación del crédito de la clientela de un banco prioriza el nivel de ingresos sobre la capacidad de ahorro como elemento indicativo de la capacidad crediticia de una persona física, esta decisión perjudicará a las mujeres en mayor medida que a los hombres como consecuencia de la brecha salarial por razón de género. Esta cuestión es relevante por cuanto la capacidad de ahorro de las mujeres es mayor que la de los hombres.

Uno de los elementos esenciales a destacar, en este sentido, es que los algoritmos, en muchas ocasiones, no discriminan con base en las denominadas categorías sospechosas, sino que lo hacen al emplear criterios de medición aparentemente neutros que, sin embargo, perjudican en mayor medida los miembros de grupos desaventajados que a los no miembros.

La discriminación algorítmica también puede venir derivada de los errores o sesgos contenidos en las bases de datos empleadas en el desarrollo de los sistemas de toma de decisiones automatizadas. Por ejemplo, una base de datos de detenciones y condenas puede contener fundamentalmente datos sobre individuos de minorías étnicas y raciales como consecuencia de la discriminación sistémica sufrida por estos grupos poblacionales tanto en sus relaciones con los cuerpos y fuerzas de seguridad como con el sistema de justicia. En este caso, el algoritmo aprenderá, de manera indirecta, que la pertenencia a una minoría étnica o racial está relacionada con un mayor índice de criminalidad.

En otros casos, el uso de algoritmos puede servir para perpetuar estereotipos en los que se basan las estructuras sociales de discriminación. Por ejemplo, no son pocos los casos en que se ha detectado que los resultados obtenidos como consecuencia de la introducción de determinadas combinaciones de palabras en los motores de búsqueda de Internet

(principalmente Google) reproducen los roles de género, cosifican a las mujeres o contribuyen a la consolidación de estereotipos negativos sobre las personas no blancas.

También existe la posibilidad de que no sea el propio algoritmo el que discrimine, sino que las decisiones que se tomen en relación con los resultados que genera sean discriminatorias.

Asimismo, es necesario tener en cuenta que los grupos poblacionales que recurren a las ayudas y servicios públicos dirigidos a mitigar la desigualdad se encuentran en situaciones de mayor vulnerabilidad económica y disponen de menos recursos para la defensa de sus derechos frente a los posibles errores cometidos en los procesos de concesión y provisión de ayudas y servicios públicos. Es por ello que la automatización de esta clase de actividad produce una serie de riesgos especialmente perjudiciales para la reducción de las diferencias sociales. Así, estos riesgos no se refieren tanto a situaciones específicas de discriminación en las que se privilegie a un subgrupo sobre otro dentro de una categoría sospechosa, sino a que los errores que pueden producir estos algoritmos contribuyen a perpetuar la posición de desventaja y vulnerabilidad de aquellos colectivos a quienes se dirigen, de manera específica, las ayudas y servicios públicos.

6. El marco jurídico europeo dirigido a proteger los derechos a la igualdad y no discriminación ofrece una serie de mecanismos que pueden ser empleados para combatir las situaciones de discriminación algorítmica

La protección frente a la discriminación desarrollada en Europa se basa en las prohibiciones generales de discriminación con base en las denominadas categorías sospechosas contenidas en la Convención Europea de Derechos Humanos y la Carta de Derechos Fundamentales de la UE. En el marco de la UE, el desarrollo específico de dichas prohibiciones se contiene en las Directivas de Igualdad que clasifican las situaciones de discriminación como discriminación directa o indirecta. Partiendo de este marco jurídico y de la interpretación del mismo llevada a cabo por el Tribunal de Justicia de la UE y el Tribunal Europeo de Derechos Humanos, según sus correspondientes competencias, puede concluirse que este sistema ofrece una base que puede ser adaptada a las necesidades de protección derivadas de la discriminación algorítmica.

6.1. La discriminación algorítmica directa puede tener lugar tanto cuando se infiere la categoría sospechosa de otros datos como cuando esta se introduce de manera explícita en el algoritmo

La discriminación directa tiene lugar cuando una persona es tratada de manera menos favorable por razón de su pertenencia a uno de los subgrupos contenidos en las categorías sospechosas. Es por ello que la discriminación algorítmica directa puede tener lugar cuando se introduzcan en el algoritmo los datos relativos a la pertenencia a un grupo desaventajado y a dicha pertenencia se asocie un valor negativo. Ese valor negativo puede conducir directamente a una consecuencia final negativa (por ejemplo, ser de procedencia hispana implica de manera automática la no concesión de un crédito) o puede contribuir a reducir la puntuación o empeorar el resultado final, pero no ser determinante para su obtención.

La discriminación directa también puede tener lugar cuando se infiera la pertenencia de una persona a un grupo desaventajado de otros datos a los que se atribuya un valor negativo. Los diseñadores del algoritmo pueden articular dichas inferencias de manera consciente o inconsciente y también es posible que el algoritmo las desarrolle una vez que se ponga en funcionamiento. Para que estos casos entren en el ámbito de la discriminación directa, los datos de los que se infiere la pertenencia al grupo especialmente protegido no pueden ser criterios aparentemente neutros, es decir, no deben ser significativos para predecir la variable objetivo. Por ejemplo, un algoritmo puede aprender que las personas que hayan pertenecido a asociaciones de mujeres no son buenas candidatas para un puesto de trabajo simplemente porque en la base de datos de la que aprende no había prácticamente mujeres y, por tanto, no había prácticamente personas pertenecientes a asociaciones de mujeres. Aunque la categoría "sexo" no aparezca en el procesamiento, el sistema la inferirá de esta otra categoría de datos que, en realidad, no predice si una persona será buena empleada o no. Igual que en la otra clase de discriminación algorítmica directa, el valor negativo asociado a la categoría de datos inferida podrá determinar de manera directa el resultado o contribuir a él, pero no ser necesariamente determinante en su obtención.

6.2. La posibilidad de probar la discriminación algorítmica directa dependerá de si el sistema incorpora la categoría sospechosa de manera explícita, si la pertenencia a un grupo desaventajado determina un resultado negativo evidente para todo el grupo y, en última instancia, de la transparencia y acceso a los datos y resultados del sistema

Los casos de discriminación algorítmica directa en los que la pertenencia al grupo explícitamente incluida determine de manera automática un resultado negativo serán relativamente fáciles de probar incluso sin acceder al contenido del algoritmo. Cabe destacar, en este sentido, que el TJUE considera que, en aquellos casos en los que el criterio formal en virtud del cual se toma una decisión sea inseparable de la pertenencia al grupo desaventajado, la acción será constitutiva de discriminación directa. Así, aunque no se disponga de la información relativa a la inclusión de la categoría sospechosa como dato que determine la decisión algorítmica, en caso de afectar esta de manera negativa a todos los miembros del grupo desaventajado, deberá considerarse como un supuesto de discriminación directa.

Sin embargo, en aquellos casos en los que se produce un caso de discriminación algorítmica directa que no afecta de manera evidente a todas las personas pertenecientes al grupo protegido, será mucho más complejo probar dicha situación. En estos casos, salvo que se pueda acceder al contenido del algoritmo, no será fácil probar la existencia de una situación de discriminación. Es más, en algunas ocasiones será incluso difícil proporcionar un elemento de comparación si los datos de los resultados obtenidos por otras personas sometidas al mismo proceso de toma de decisión se encuentran protegidos por la normativa en materia de protección de datos. Asimismo, con respecto a la aportación de un elemento de comparación, dado que los algoritmos que se utilizan de manera creciente tienen la capacidad de aprender y desarrollarse de manera autónoma, puede suceder que sea casi imposible encontrar supuestos la decisión en relación con los cuales haya sido adoptada sobre exactamente los mismos parámetros.

Para aquellos casos en los que no se dé acceso pleno al algoritmo, deberá establecerse la posibilidad de experimentar con el sistema, introduciendo diferentes perfiles para comprobar si este efectivamente discrimina o no con base en alguna de las categorías sospechosas. Ahora bien, si el sistema no tiene en cuenta las categorías sospechosas sino que las infiere de otros datos, será prácticamente imposible deducir si se trata de un supuesto de discriminación directa o indirecta ya que, sin acceso al sistema, no será posible determinar si las variables que se tienen en cuenta, y que derivan en un trato discriminatorio, tienen valor predictivo o no para el objetivo para el que se utiliza el algoritmo.

6.3. Los supuestos de discriminación algorítmica indirecta podrán probarse empleando, sobre todo, métodos estadísticos, pero las exigencias con respecto a la cantidad de personas afectadas del grupo deberían rebajarse en estos casos

El enjuiciamiento de los supuestos de discriminación indirecta requiere que se acredite la existencia de una disposición, práctica o criterio que resulte en efectos más perjudiciales para las personas pertenecientes al grupo protegido que a las no pertenecientes a dicho grupo. Una vez que se acredita la existencia, a primera vista, de un supuesto de discriminación indirecta como el descrito, se traslada la carga de la prueba a la parte demandada, que habrá de demostrar que no se ha vulnerado el principio de igualdad de trato.

En el caso de la discriminación algorítmica, la disposición, práctica o criterio aparentemente neutral puede ser la variable específica introducida en el sistema que, por el valor que le otorga este, genera un resultado discriminatorio para el grupo desaventajado, o bien el algoritmo generalmente considerado. El primer caso será más complicado de establecer puesto que para determinar la variable o variables que generan el resultado discriminatorio se necesitará tener acceso al sistema. Es más, en muchas ocasiones, la gran cantidad de datos procesados y la complejidad de los sistemas de procesamiento impiden identificar de manera clara cuáles son los elementos que condicionan los resultados del sistema. Es por ello que, como norma general, será el sistema en su totalidad el que deba ser considerado como disposición, práctica o criterio aparentemente neutral que genera resultados discriminatorios.

Igual que en el caso de la discriminación directa, la prueba de la discriminación indirecta requiere de la aportación de un elemento de comparación. Puesto que la figura de la discriminación indirecta se centra en el resultado y no tanto en el trato, en ocasiones, tanto el Tribunal de Justicia de la UE como el Tribunal Europeo de Derechos Humanos han basado la existencia a primera vista de un caso de discriminación indirecta en datos estadísticos.

Los casos en que se enjuicien las decisiones tomadas por sistemas automatizados, sin duda, comportarán un aumento de la incidencia de la prueba basada en datos estadísticos. El objetivo de estos sistemas es analizar personas en masa y continuamente dar resultados sobre una misma cuestión. Por eso, en principio, será posible disponer de una importante cantidad de ejemplos que aportar como comparación.

Ni el TJUE ni el TEDH han fijado un criterio único sobre qué porcentaje de personas negativamente afectadas por la disposición, criterio o práctica aparentemente neutra debe pertenecer al grupo desaventajado o el porcentaje de personas del grupo que deben resultar negativamente afectadas por la medida en comparación con el porcentaje de personas no pertenecientes para considerar que se da un caso de discriminación directa. Sin embargo, el

TJUE sí ha indicado que la diferencia debe ser considerable y que el resultado perjudicial debe tener una incidencia de en torno al 80 o 90% para las personas pertenecientes al grupo protegido.

En todo caso, tanto el TJUE como el TEDH tienden a analizar cada situación de manera específica. Así, en el caso de *D. H. y otros contra la República Checa*,[1726] el TEDH consideró que la desproporcionada presencia de menores romaníes en centros escolares para niñas y niños con necesidades especiales constituía una medida indirectamente discriminatoria. En este caso, la presencia de menores romaníes en estas escuelas se encontraba en torno a un 50%. Sin embargo, el tribunal tuvo en cuenta dicho dato en relación con la composición de la población del país, pues únicamente un 2% de las y los menores checos eran romaníes.

Por su parte, la sentencia dictada en el caso *Seymour-Smith*[1727] abre una importante puerta a la reducción de los límites mínimos para considerar la prueba estadística válida. En dicho supuesto, se estableció la posibilidad de apreciar la concurrencia de discriminación "si los datos estadísticos mostraran una diferencia menos importante, pero persistente y relativamente constante durante un largo período de tiempo", en aquel caso, entre mujeres y hombres trabajadores. Teniendo en cuenta que los algoritmos son generalmente implementados con el objetivo de automatizar procesos que se reproducen una y otra vez, será relativamente fácil obtener una serie longitudinal de resultados producidos por estos sistemas. Es más, una vez que un algoritmo incorpora un sesgo que perjudica a las personas pertenecientes a un grupo desaventajado, salvo que el sistema incorpore instrucciones para corregir estos sesgos, algo que generalmente no sucede, lo habitual es que dicho sesgo se perpetúe durante todo el periodo de funcionamiento del sistema.

Ahora bien, como ya se ha indicado en el caso de la discriminación algorítmica directa, será necesario establecer controles y mecanismos que den validez a los elementos de comparación y series de resultados longitudinales aportados por algoritmos de autoaprendizaje cuyo contenido, razonamiento e instrucciones pueden ir variando.

La prueba estadística de la discriminación algorítmica puede adoptar diversas formas. Por ejemplo, como en los casos ya estudiados se puede fijar como una medida de la paridad en el grupo. Es decir, de todas las personas candidatas para un puesto de trabajo, ¿a cuántas

[1726] Sentencia TEDH, D.H. y otros c. la República Checa, 57325/00, 13 de Noviembre de 2007.
[1727] Sentencia TJUE 9 de febrero de 1997, C-167/97, Regina contra Secretary of State for Employment, *ex parte* Nicole Seymour-Smith y Laura Perez, párrafo 60.

mujeres selecciona y qué porcentaje de las candidaturas representan? O, de todas las mujeres que se postularon, ¿a qué porcentaje se seleccionó? En este segundo caso cabría comparar dicho resultado con el porcentaje de hombres seleccionados del total que presentó su candidatura.

Ahora bien, estas medidas no son siempre adecuadas, puesto que se centran en el resultado final absoluto y hay determinados ámbitos en los que otros principios, intereses o derechos deberán primar sobre la igualdad. Por ejemplo, si comparamos la presencia de hombres no blancos y hombres extranjeros en las cárceles en relación con la presencia de hombres blancos nacionales y encontramos que la presencia de los primeros es desproporcionada en relación con el porcentaje de la población que constituyen, podremos detectar que existe un problema. Este problema se puede situar en la discriminación estructural e institucional presente en los sistemas de fuerzas del orden, judicial y penitenciario o en las situaciones de exclusión social y experiencias que suelen compartir las personas migradas y no blancas. Sin embargo, no se puede considerar que un sistema penitenciario es indirectamente discriminatorio siempre que no sea absolutamente representativo de los diferentes segmentos poblacionales. Ello comportaría una construcción e implicaciones absurdas de las nociones de acción afirmativa y de discriminación indirecta.

Es por ello que en muchos supuestos de discriminación algorítmica indirecta será más conveniente recurrir a medidas de paridad en la precisión. Esto es, si el algoritmo predice de manera igualmente precisa que personas blancas y no blancas reincidirán, dicho algoritmo no será discriminatorio. Ahora bien, surge un nuevo problema a la hora de determinar cuál sería la medida a emplear para determinar si un algoritmo no cumple con los criterios de paridad en la precisión. Por ejemplo, se podría determinar a través de las tasas de falsos positivos y negativos. En este caso, deberíamos preguntar si la tasa de falsos positivos y falsos negativos es la misma para los diferentes grupos. Otra posibilidad sería calcular si la tasa de error en el cálculo de positivos y negativos de los diferentes grupos se desvía significativamente. También se podría calcular como la relación de falsos positivos a falsos negativos y la relación de falsos negativos a falsos positivos para cada grupo. Por último, la medida escogida podría calcular la cantidad total de verdaderos positivos y negativos como porcentaje dentro del propio grupo.

No existe una única respuesta de cuál haya de ser la medida más adecuada para determinar la precisión paritaria del algoritmo pero todas ellas ofrecen un elemento probatorio adicional en

los casos de discriminación indirecta y, por ello, la posibilidad de examinar la igualdad y no discriminación en ámbitos en los que no basta con demostrar la desigualdad en números absolutos.

6.4. De no articularse con sumo cuidado, la posibilidad de justificar la discriminación indirecta podría abrir la puerta a considerar lícita toda forma de discriminación algorítmica indirecta

La justificación de la discriminación indirecta pasa, en primer lugar, por la prueba de que la medida discriminatoria persigue un fin legítimo. En general, el TJUE ha dado por válidos prácticamente todos aquellos fines siempre que el propio fin constituya discriminar o no esconda una intención de discriminar de manera evidente. Sí cabe hacer una especial referencia a la consideración de lo que constituyen fines legítimos para el sector público puesto que el TJUE ha determinado que las consideraciones de carácter presupuestario no pueden, por sí solas, suponer un fin legítimo que ampare una medida indirectamente discriminatoria.[1728] Por consiguiente, y aun teniendo en cuenta las mayores exigencias impuestas a los poderes públicos en la justificación de los fines de las medidas discriminatorias, es probable que se considere legítima, cualquier finalidad (lícita) para la que se emplee un algoritmo.

Una vez determinada la concurrencia de una finalidad legítima, los tribunales deben proceder a realizar el test de proporcionalidad de la medida discriminatoria que incorpora la evaluación de la adecuación de la medida a los fines que persigue; su necesidad, esto es, la inexistencia de otra medida menos gravosa que alcance los mismos fines y, por último, el análisis de proporcionalidad en sentido estricto.

La adecuación del algoritmo discriminatorio a los fines que persigue podrá establecerse, en los casos en que se tenga acceso al contenido del algoritmo, a través de la determinación del significado predictivo (o falta de este) de aquellas variables que conducen a resultados discriminatorios. En caso de que se determine que dichas variables no son útiles para predecir la variable objetivo y que, sin embargo, producen resultados discriminatorios, el algoritmo deberá, automáticamente clasificarse como un sistema que produce discriminación directa.

[1728] Sentencias TJUE de 11 de noviembre de 2014, C-530/13, Leopold Schmitzer contra Bundesministerin für Inneres, párrafo 41 y de 21 de julio de 2011, Asuntos Acumulados C-159/10 y C-160/10, Gerhard Fuchs (C-159/10), Peter Köhler (C-160/10) contra Land Hessen, párrafo 74.

Ahora bien, teniendo en cuenta la creciente complejidad y opacidad de los sistemas automatizados, es poco probable que puedan llegar a identificarse aquellas variables que producen resultados discriminatorios. También podría determinarse la concurrencia de discriminación algorítmica a través de sesgos en la base de datos, debiendo la parte demandada probar si, pese a dichos sesgos, la base de datos es adecuada para predecir la variable objetivo del algoritmo.

En todo caso, es más probable que el test de adecuación se realice sobre todo el sistema en su conjunto, debiendo responder a la pregunta: ¿es el algoritmo adecuado para predecir el fenómeno del que se ocupa? La adecuación deberá, pues, medirse a partir de la precisión predictiva del sistema.

Con respecto a la necesidad de utilización del sistema discriminatorio, la parte demandada deberá probar que valoró otras posibilidades y mecanismos para la consecución del mismo fin para el que se emplea el algoritmo. Es, sobre todo, de vital importancia que no se acepte una justificación genérica de la necesidad basada en la idea de que, como norma, los algoritmos son menos discriminatorios que los seres humanos y, por tanto, el uso de un sistema automatizado de toma de decisiones implica, de manera automática, la superación del test de necesidad.

Existe una creciente rama del conocimiento dedicada a la creación de sistemas automatizados no discriminatorios. Es por ello que la parte demandada deberá acreditar que no existía un algoritmo menos discriminatorio ni una base de datos más completa y menos sesgada. Ahora bien, con el objetivo de no imponer costes inasumibles a algunos operadores tecnológicos, deberá exigírseles el uso de sistemas no discriminatorios dentro del criterio de "las mejores técnicas disponibles". Es decir, deberá tenerse en cuenta tanto el estado de la técnica como la capacidad económica de la parte demandada. En este sentido, como ya se expuso en el último capítulo de este trabajo y volverá a indicarse en el último punto de las conclusiones, debería instituirse un marco europeo de "mejores técnicas disponibles" aplicable al uso de tecnologías de procesamiento de datos y de toma de decisiones automatizadas.

Finalmente, con respecto al test de proporcionalidad en sentido estricto, será sobre todo complicado determinar si, en aquellos casos que podrían englobarse en la categoría de "discriminación estadística precisa", debe prevalecer la igualdad sobre la libertad y la eficiencia como derechos y/o intereses de la parte demandada. Sobre todo, debe tenerse en

cuenta que, incluso en aquellos casos en que el resultado que perjudica a las personas pertenecientes a grupos desaventajados sea realmente preciso, es decir, que refleja de manera adecuada la realidad, dicha realidad es el producto de la construcción de la sociedad sobre una serie de estructuras de discriminación. Para la realización de este último examen de proporcionalidad proponemos que se lleve a cabo un análisis de costes y beneficios, evaluando, en el contexto analizado, todos los bienes y derechos en juego. En dicho examen deberá, en todo caso, tenerse en cuenta la capacidad humana para comprender determinados elementos intangibles que no son capturados por las herramientas informáticas.

Asimismo, también cuando la utilización de determinados datos que revelen la pertenencia a un grupo protegido sea necesaria, por ejemplo respecto del análisis de la situación económica de una persona para determinar si se le concede un préstamo o no, deberá comprobarse que el resultado desigual entre una persona perteneciente a una clase socioeconómica baja y una persona con mayor capacidad económica no debe ser desproporcionado. Esto es así, sobre todo si se tiene en cuenta que existe una relación directa entre la pertenencia a determinados grupos protegidos, como minorías étnicas, y la pertenencia a estratos socioeconómicos más bajos.

Asimismo, consideramos conveniente que se impongan mayores requisitos en el respeto a los derechos a la igualdad y no discriminación a las grandes compañías tecnológicas, tanto por su poder económico como por su gran presencia en el mercado, con comportamientos que pueden definirse como monopolísticos, y por la gran influencia que sus sistemas tienen sobre la esfera de la dignidad humana y los derechos fundamentales de las personas.

6.5.Discriminación algorítmica por asociación

La figura de la discriminación por asociación, reconocida por el TJUE en los casos *Coleman*[1729] y *Chez*[1730]*,* abre la posibilidad de reconocer como discriminación algorítmica (directa o indirecta) aquellos supuestos en los que el algoritmo no seleccione a las personas por su pertenencia a un grupo protegido sino por su relación con personas del grupo. En este sentido, el sistema empleado por Facebook para excluir a personas relacionadas con determinados grupos raciales de anuncios de viviendas entraría dentro de esta categoría, ofreciendo tanto a las personas pertenecientes al grupo como a las no pertenecientes pero de

[1729] Sentencia TJUE de 17 de julio de 2008, C-303/06, S. Coleman v. Attridge Law and Steve Law.
[1730] Sentencia TJUE de 16 de julio de 2015, C-83/14, CHEZ Razpredelenie Bulgaria AD v. Komisia za zashtita ot diskriminatsia.

alguna manera vinculadas al mismo, la especial protección ofrecida a los derechos a la igualdad y no discriminación cuando estos operan en relación con las categorías sospechosas. Asimismo, este mecanismo ofrece una vía para proteger a aquellas personas que, por ejemplo, no quieran hacer pública su orientación sexual al considerarlas como parte del grupo a proteger por estar relacionadas con aquel y, no, necesariamente, por efectivamente pertenecer al mismo.

6.6. Marco jurídico de igualdad sustantiva

Las Directivas Europeas de Igualdad también contienen una serie de preceptos dirigidos a la construcción de un marco de igualdad sustantiva, reconociendo, por ejemplo, la posibilidad de implementar acciones positivas. El fomento de la igualdad entre mujeres y hombres como objetivo general también se encuentra contenido en el Tratado de Funcionamiento de la UE. Las principales medidas para la realización efectiva del marco jurídico de igualdad sustantiva son las acciones afirmativas, las medidas de fomento de igualdad y el desarrollo de acciones para la generalización de la transversalidad de la igualdad.

En general, el TJUE se ha mostrado reacio a aceptar la introducción de medidas de acción positiva. Por ejemplo, solo ha aceptado que se diese prioridad al grupo desaventajado cuando se introduzca una cláusula que permita considerar las características específicas concurrentes en las personas pertenecientes al grupo aventajado, otorgando cierta flexibilidad al mandato de priorizar la selección de mujeres.

La igualdad sustantiva algorítmica ha de construirse sobre la noción de "igualdad en el diseño", es decir, deben considerarse los distintos elementos que pueden derivar en situaciones de discriminación algorítmica o en usos que contribuyan a perpetuar determinadas estructuras de desigualdad en cada momento de la construcción y desarrollo de los sistemas algorítmicos, y también después de su puesta en funcionamiento. Partiendo de esta premisa, existen varias propuestas relativas al nivel de igualdad al que deberían aspirar los sistemas algorítmicos, así como sobre los mecanismos existentes para su consecución.

Aquellos mecanismos que persiguen una igualdad de resultados total, basados en nociones de paridad estadística, difícilmente podrán superar el examen de aceptabilidad impuesto por el TJUE. Ahora bien, la introducción de determinados elementos que flexibilicen los mandatos de acción afirmativa sí podrían facilitar la aceptación de la acción afirmativa algorítmica. En

este sentido, se han puesto de relieve sendas propuestas dirigidas a introducir acciones afirmativas en el desarrollo de algoritmos cuyo objetivo es evitar generar grandes distorsiones con respecto a los resultados que serían obtenidos sin las modificaciones tendentes a crear algoritmos más igualitarios.

En cualquier caso, si bien resulta relevante la introducción de acciones afirmativas para compensar la desigualdad estructural sufrida por determinados grupos, también es de gran relevancia que se acuda a la raíz del problema y se examinen las bases de datos empleadas en la creación y desarrollo de los algoritmos; las decisiones previas sobre las que pueden aprender estos sistemas y, las variables y el peso relativo otorgado a cada una de ellas pues, en muchas ocasiones, los resultados discriminatorios encuentran su origen en la construcción de la sociedad sobre unas estructuras que dan prioridad a los valores y características más comunes en los miembros de grupos que, históricamente, han ocupado el poder.

Finalmente, también pueden y deben emplearse los sistemas automatizados para detectar actitudes, estructuras y situaciones discriminatorias de manera más efectiva con el objetivo de mejorar las actuaciones en materia de lucha contra la discriminación y promoción de la igualdad.

7. El marco jurídico europeo en materia de igualdad y no discriminación adolece de una serie de deficiencias que, en ocasiones, son comunes a la protección frente a situaciones de discriminación algorítmica y decisiones de discriminación llevadas a cabo sin la mediación de sistemas automatizados y, en otras, afectan de manera específica a la discriminación algorítmica

El marco jurídico europeo en materia de igualdad confiere una protección más específica y extensa en aquellos supuestos previstos en las Directivas de Igualdad. Si bien es cierto que la Carta Europea de Derechos Fundamentales establece una serie de obligaciones a los Estados y, también, hasta cierto punto, a los particulares, en materia de respeto y protección a los derechos a la igualdad y la no discriminación, lo cierto es que el artículo 21 de la Carta deja un considerable margen de actuación en la determinación de las situaciones y forma de protección de estos derechos. Sin embargo, las Directivas de Igualdad incorporan un marco mucho más completo, definiendo, de manera más extensa, no solo aquellas categorías sospechosas, sino también las clases de discriminación y los ámbitos en los que los derechos a la igualdad y a la no discriminación merecen una especial protección. Las Directivas de

igualdad constituyen el límite mínimo de protección que los Estados miembros de la UE deben respetar en sus ordenamientos jurídicos internos.

Es por ello que la desigual protección ofrecida por estas Directivas genera importantes limitaciones en la protección de la igualdad y la no discriminación de la ciudadanía de la Unión. Cabe destacar que la prohibición de discriminación en el acceso a bienes y servicios se aplica, únicamente, por razón de raza o género y no así por razón de las restantes categorías sospechosas. Asimismo, no se incluye la publicidad en el ámbito de aplicación de la Directiva relativa a la igualdad de hombres y mujeres en el acceso a bienes y servicios. Por su parte, la protección frente a la discriminación por razón de religión, convicciones, edad, orientación sexual y discapacidad únicamente tiene lugar en el ámbito del empleo.

Se limita así la posibilidad de reconducir, por la vía de la publicidad discriminatoria, la perpetuación de estereotipos de género a través de los algoritmos empleados por los motores de búsqueda (principalmente, Google) o la utilización de algoritmos que impidan mostrar publicidad relativa a determinados productos o servicios a cualquier persona perteneciente a un grupo desaventajado, salvo en los casos en los que la exclusión se realice por razón de raza. Asimismo, teniendo en cuenta la especial vulnerabilidad de las personas pertenecientes a estratos socioeconómicos más bajos frente al creciente uso de sistemas automatizados, también resulta especialmente problemático que la clase social no se incluya como categoría especialmente protegida ni en la Convención Europea de Derechos Humanos ni en la Carta de Derechos Fundamentales de la UE y, por supuesto, tampoco en las Directivas.

Entre las deficiencias del sistema europeo de protección de la igualdad y la no discriminación también cabe destacar, sobre todo con referencia al uso de algoritmos, las dificultades derivadas de la necesidad de aportar un elemento de comparación. Por una parte, cabría esperar que la cantidad de situaciones procesadas por un mismo algoritmo pudiera dar lugar a suficientes elementos de comparación. Ahora bien, si los datos referentes a los restantes resultados producidos por el algoritmo se encuentran protegidos por la normativa en materia de protección de datos será más costoso, aunque no imposible, acceder a dichos elementos de comparación. Así, la razón por la que no es imposible es porque, o bien pueden solicitarse dichos datos una vez sean anonimizados, o se pueden realizar experimentos con el programa, introduciendo diferentes perfiles para comparar los resultados entre unos y otros.

Ahora bien, también puede suceder que la constante evolución de los sistemas derive en que no existan dos casos en los que el algoritmo haya funcionado de manera idéntica. En este caso, será más difícil recurrir a la solución de la experimentación con el algoritmo. Es por ello que resulta necesario que los encargados y responsables del procesamiento de datos tengan y cumplan con la obligación de guardar una copia de cada versión del programa.

También cabe destacar que si bien es cierto que la posibilidad de emplear pruebas estadísticas en los supuestos de discriminación aumente las posibilidades de acreditar la existencia de vulneraciones a los derechos a la igualdad y a la no discriminación, la dificultad de comprender el contenido de los sistemas también impide que el propio sistema se pueda considerar discriminatorio de plano, sin atender a elementos de comparación. Por ejemplo, en un supuesto en el que había una única mujer candidata a un ascenso y fue rechazada por estar embarazada, el TJUE consideró que se trataba de un supuesto de discriminación directa por razón de sexo puesto que el rechazo a una persona para un puesto de trabajo por el hecho de estar embarazada solo se puede dar en los casos en que la persona candidata sea mujer.[1731] Esta consideración difícilmente podrá darse en un caso en que la decisión la tome un algoritmo a cuyo contenido no se pueda acceder.

Asimismo, la falta de consideración de la discriminación estructural como elemento de análisis en los casos de discriminación también supone una importante limitación en la protección efectiva de los derechos a la igualdad y no discriminación. Esta limitación resulta especialmente relevante en el caso de la discriminación algorítmica por cuanto la discriminación estructural, como categoría analítica permite entender que, incluso en los casos en los que la discriminación es aparentemente precisa, esta es el resultado de una sociedad que, de manera sistemática, sitúa a las personas pertenecientes a determinados grupos en una posición de desventaja. Por tanto, si no se corrigen estas situaciones de discriminación aparentemente eficiente, no se corregirán las estructuras que sitúan en un punto de partida desigual a las personas pertenecientes a grupos desaventajados.

Cabe también destacar la poca consideración otorgada a los supuestos de discriminación interseccional, también conocida como discriminación múltiple. Las personas pertenecientes a más de un grupo desaventajado pueden ser y son discriminadas por razones específicas que se pueden diferenciar de aquellas situaciones de discriminación basadas en una única

[1731] Sentencia TJUE de 8 de Noviembre de 1990, C-177/88, Elisabeth Johanna Pacifica Dekker v. Stichting Vormingscentrum voor Jong Volwassenen (VJV Centrum) Plus, párrafo 12.

categoría especial. En este sentido, la discriminación sufrida por una mujer migrada no es equivalente ni a la discriminación sufrida por los hombres migrados ni a la discriminación sufrida por las mujeres nacionales de un Estado. La necesidad de prestar atención a esta clase de discriminación deviene, si cabe, más acuciante como consecuencia del creciente uso de sistemas automatizados. La cantidad de datos cruzados por los sistemas algorítmicos implica un aumento de la personalización de los resultados y, por tanto, un aumento de las respuestas específicas dadas a aquellas personas en las que concurran varias categorías que las hagan especialmente vulnerables. Es por ello que la realidad de la discriminación interseccional y la forma en la que esta se refleja en la discriminación algorítmica, debe ser abordada. En este sentido, el TEDH ha realizado una importante labor en el reconocimiento progresivo de las situaciones de discriminación que afectan de manera específica a aquellas personas sobre las cuales se toman decisiones con base en la concurrencia en ellas de dos o más categorías sospechosas.

Por último, el sistema europeo de reconocimiento y protección de los derechos a la igualdad y a la no discriminación adolece de importantes deficiencias que dificultan la efectiva aplicación de los mecanismos tendentes a salvaguardar los derechos que nos ocupan. En este sentido, cabe destacar que, si bien existe una red de instituciones dedicadas a la protección de la igualdad, estas se ocupan, principalmente, de realizar actuaciones de investigación y no siempre tienen suficiente capacidad para controlar el respeto a los derechos a la igualdad y a la no discriminación.

Asimismo, en muchas ocasiones el derecho de acceso a la justicia se ve condicionado tanto por las propias limitaciones derivadas de los costes económicos y burocrático-administrativos de los procesos de protección de los derechos a la igualdad y no discriminación como por el hecho de que las víctimas de las situaciones de discriminación en muchas ocasiones no son tan siquiera conscientes de que sus derechos están siendo vulnerados. Por ejemplo, la víctima de discriminación en un proceso de selección de personal llevado a cabo por un algoritmo, o a quien se le ha denegado un préstamo por sistema automatizado que basa su decisión en una categoría sospechosa, difícilmente serán conscientes de las razones verdaderas por las que no se les contrata o se les deniega un préstamo, especialmente si no tienen acceso o conocimiento de otras personas que hayan participado en el mismo proceso.

A mayor abundamiento, la práctica inexistencia de la adopción de acciones colectivas en materia de igualdad y no discriminación en Europa también supone un importante

impedimento a la efectiva protección de estos derechos, puesto que estas acciones otorgan a las víctimas de discriminación una red y mecanismos de apoyo que facilitan la realización de todos los trámites burocráticos y reducen el coste económico de la defensa de sus derechos.

8. El fenómeno de la discriminación algorítmica y sus posibles soluciones jurídicas guardan una íntima conexión y han de estudiarse en relación con los restantes riesgos generados por el creciente uso de los sistemas automatizados de toma de decisiones

La creciente capacidad computacional de las tecnologías de procesamiento de datos y de automatización de decisiones genera una serie de riesgos para los derechos fundamentales de las personas, así como para los valores propios de los Estados democráticos de Derecho, que no se limitan, exclusivamente, a la esfera de la igualdad y no discriminación.

En primer lugar, estos sistemas no solo discriminan siguiendo los mandatos implícitos de las estructuras tradicionales de subordinación, sino que también cometen errores o toman decisiones no ajustadas a la realidad de las que se derivan importantes perjuicios para las personas afectadas. También, con respecto a la forma en la que estos sistemas afectan a la esfera inherente a los derechos fundamentales de la persona, cabe destacar que la extracción de datos personales y creación de perfiles empleados en la predicción de los comportamientos de las personas supone un ataque a su dignidad, autonomía e individualidad así como, y de manera más evidente, a su derecho a la intimidad y a la protección de datos de carácter personal.

La cantidad de datos recabados que se van incorporando a los perfiles digitales de las personas físicas ayudan a desarrollar un "doble digital" sobre el que el individuo tiene poca o nula capacidad de control y que, sin embargo, puede llegar a influir de manera considerable en la toma de decisiones, tanto por parte de sujetos privados como de los poderes públicos, con una directa incidencia en la esfera de sus derechos e intereses. Este o esta "doble digital" se crea comparando a un ser humano con otras personas similares y no atendiendo a sus especificidades y particulares vivencias como individuo. Es por ello que la persona física pierde su individualidad para convertirse en una parte del rebaño cuyas preferencias y características se miden, comparan y analizan, constantemente y cada vez con más capacidad de realizar inferencias a partir de ello, en relación con las del resto de la sociedad o grupo de contraste. Cabe destacar que las personas pertenecientes a colectivos desaventajados tienden

a verse especialmente perjudicadas cuando se operan estos análisis, dado que no se suelen tener en cuenta aquellas tradiciones o características propias del grupo que les pueden beneficiar en los procesos de toma de decisiones. Asimismo, estos sistemas están en ocasiones diseñados con el objetivo de manipular a las personas, limitando su autonomía y libertad o cuando menos de influir en su comportamiento con la orientación que más conviene a quienes los diseñan y utilizan.

Todo lo expuesto en este párrafo conlleva importantes afecciones a la esfera de la dignidad de las personas que, en muchas ocasiones, son tratadas como medios para conseguir un fin diferente y ajeno a su voluntad, realidad que se acentúa si consideramos la dificultad que en ocasione se da de acceder a procesos de reevaluación de las decisiones tomadas en los que intervengan seres humanos. Así, en segundo lugar y en relación con la última afirmación, la incorporación de estos sistemas a diferentes procesos no siempre ha venido acompañada de una correlativa introducción de derechos de información y defensa de las personas afectadas por ellos. Esto quiere decir que cuando se cometen errores o se toman decisiones que no están correctamente justificadas, las personas afectadas no disponen de los medios tradicionales para informarse y recurrir la decisión adoptada. Esta realidad se ha puesto sobre todo de relieve en el marco del uso por los poderes públicos de sistemas algorítmicos en la toma de decisiones. Así, podemos encontrar diferentes ejemplos en los que se ha dado una respuesta negativa a la solicitud de acceso a la información del algoritmo. Es por ello que una parte importante de la doctrina ha resaltado la necesidad de reconocer una serie de derechos al debido proceso tecnológico.

En tercer lugar, cabe destacar la opacidad de estos sistemas como consecuencia de su complejidad y, en ocasiones, también de la falta de voluntad de revelar su contenido por parte de los responsables y encargados del tratamiento de datos. La falta de transparencia o comprensibilidad de estos sistemas dificulta enormemente que las personas destinatarias de las decisiones, sean estas discriminatorias o no, puedan ser realmente conocedoras de la justificación subyacente al resultado del procesamiento y, por tanto, que puedan articular su defensa de manera efectiva. Asimismo, la cantidad de actores públicos y privados que participan en la creación de algunos de estos sistemas también obstaculiza la identificación de las personas responsables de aquellas decisiones que puedan generar daños que deban ser resarcidos.

Por último, los riesgos derivados de la creciente automatización de la toma de decisiones en cada vez más ámbitos generan una serie de problemas específicos, y de mayor gravedad, cuando nos referimos a actuaciones de los poderes públicos, ya que sus actuaciones deben regirse, en mayor medida que en el sector privado, por una serie de reglas y principios como las obligaciones de transparencia, la necesaria justificación de sus decisiones y el respeto a una serie de procedimientos predeterminados por la ley. Asimismo, la incorporación de sistemas automatizados a las actuaciones de los poderes públicos en ocasiones supone una delegación de tareas que son, en principio, inherentemente públicas, sin la inclusión de los necesarios procesos de revisión y control posteriores.

9. El marco jurídico en materia de protección de datos constituye, por ahora, la principal herramienta jurídica dirigida a prevenir y lidiar, de manera conjunta, con los riesgos generados por los algoritmos y ofrece una serie de mecanismos útiles, pero también presenta importantes deficiencias, de cara a la protección frente a los posibles daños causados por las herramientas de procesamiento de datos y de toma de decisiones automatizadas

Hasta la fecha, el marco jurídico específicamente dirigido a lidiar y controlar los riesgos generados por los sistemas de procesamiento de datos y de toma de decisiones automatizadas viene dado por la normativa en materia de protección de datos. El RGPD y, en menor medida, la Directiva 2016/680, establecen un sistema que puede denominarse de gobernanza colaborativa o co-regulación, en el que se incluyen preceptos estructurados como mandatos y otros como meras recomendaciones o que precisan la cooperación de los responsables y encargados del tratamiento de datos para su efectividad.

9.1. El sistema europeo en materia de protección de datos se encuentra actualmente llamado a abordar, de manera conjunta, los diferentes riesgos derivados del creciente uso de algoritmos y de la automatización de todo tipo de procesos

Es lógico que el marco jurídico dirigido a proteger frente a los riesgos causados por las tecnologías de procesamiento de datos y la automatización de procesos parta de la protección de datos personales, estrechamente relacionado con la protección de la intimidad. Los daños causados por estos sistemas son, principalmente, el resultado de la recogida de datos de carácter personal y muchos de sus efectos tienen lugar en la esfera personal de aquellos individuos cuyos datos han sido recogidos y procesados.

También por esta razón, podría caerse en el error de considerar que las reglas y normas establecidas por el ordenamiento jurídico privado deberían ser suficientes para gestionar los conflictos surgidos en relación con el procesamiento de datos, la creación de perfiles y la toma de decisiones automatizadas. Sin embargo, sin entrar a valorar el uso de estos sistemas por los poderes públicos, también cuando los responsables y encargados del tratamiento operan en el sector privado, los daños causados y riesgos que se originan en el ámbito administrativo y cuando la actividad es realizada por los poderes públicos tienen efectos más allá de la configuración tradicional de relaciones jurídicas entre personas privadas. Así, por ejemplo, en ocasiones, los individuos afectados por el procesamiento de datos no han compartido sus datos personales sino que, por ejemplo, dichos datos han sido recopilados a través de los perfiles o cuentas de terceras personas. O, incluso, cuando hay posibilidades de ejercicio de poder público, son el resultado del cruce de bases de datos que las Administraciones públicas tienen sobre los ciudadanos y que han sido recopilados en el ejercicio de funciones públicas.

Igualmente, cabe destacar la existencia de una serie de fallos de mercado, incluyendo externalidades negativas, regímenes monopolísticos o situaciones de asimetría informativa generadas, desde una perspectiva económica, por las relaciones jurídicas articuladas en relación con el procesamiento de datos y automatización de decisiones. Todo ello contribuye a justificar la intervención pública en el mercado privado de los servicios de procesamiento de datos. Es por ello que el marco jurídico europeo en materia de protección de datos establece un sistema mixto que fija importantes limitaciones a la actuación de poderes públicos y organizaciones privadas en el procesamiento de datos, creación de perfiles y toma de decisiones automatizadas pero que también, como ya se ha indicado, cuenta con mecanismos orientados a fomentar la auto-regulación privada de esta clase de actuaciones.

El sistema europeo en materia de protección de datos contiene así tres tipos de mecanismos: i) prohibiciones de procesamiento, i) una serie de mecanismos dirigidos a establecer un proceso tecnológico justo y iii) una serie de instrumentos dirigidos al control general de los sistemas algorítmicos.

En primer lugar, las prohibiciones de procesamiento se refieren, por una parte, al procesamiento de categorías especiales de datos (arts. 9 RGPD y 10 Directiva 2016/680) y, por otra, a la toma de decisiones basadas únicamente en un tratamiento automatizado, incluida la elaboración de perfiles (arts. 22 RGPD y 11 Directiva 2016/680). Sobre todo con

respecto a la prohibición referida al procesamiento de categorías especiales de datos personales cabe referirse a que es este precepto el que sitúa principalmente al marco jurídico europeo en materia de protección de datos como un instrumento de prevención de la discriminación a través de la anti-clasificación. Así, teniendo en cuenta que las categorías especiales de datos personales coinciden, en lo sustancial, con las categorías sospechosas contenidas en los instrumentos jurídicos de protección de la igualdad y no discriminación, cabe concluir que el legislador europeo lo que pretende con este precepto, que ya se encontraba en la anterior Directiva de Protección de Datos 95/46/CE, es evitar que se tomen decisiones basadas en las categorías especialmente protegidas por el marco jurídico europeo de protección a los derechos a la igualdad y a la no discriminación, como la raza o la religión.

En segundo lugar, el sistema de proceso tecnológico justo, o debido proceso tecnológico, incluido en el ordenamiento jurídico europeo en materia de protección de datos incorpora una serie de derechos individuales de información, a ser oída u oído y a recurrir los perfiles creados y/o decisiones tomadas y obligaciones correlativas para los responsables y encargados del tratamiento de datos personales.

Finalmente, la última parte del marco jurídico en materia de protección de datos la constituyen aquellos instrumentos dirigidos a controlar, de manera general, los sistemas automatizados. Estos mecanismos incluyen la obligación de realizar evaluaciones de impacto especialmente cuando estos sistemas generen riesgos significativos para las libertades y derechos fundamentales de las personas (arts. 35 RGPD y 27 Directiva 2016/680) y un sistema de sanciones (arts. 84 RGPD y 57 Directiva 2016/680) y entidades públicas independientes de supervisión y control de protección de datos (Capítulo VI RGPD y Directiva 2016/680). Adicionalmente, como ya se ha indicado, el RGPD (y no así la Directiva 2016/680 por ir dirigida exclusivamente al sector público) incluye diversos mecanismos que requieren de la cooperación entre poderes públicos y actores privados para su efectividad, como por ejemplo, la elaboración de códigos de conducta para el procesamiento de datos (arts. 40 y 41 RGPD) y el desarrollo de mecanismos de certificación (art. 42 RGPD).

9.2. El sistema europeo en materia de protección de datos no consigue proteger, de manera efectiva, contra los numerosos riesgos surgidos del creciente uso de sistemas algorítmicos en los procesos de toma de decisiones

Las deficiencias del marco jurídico europeo en materia de protección de datos tienen su origen en las insuficiencias generadas por un sistema que pretende proteger frente a una importante y diversa cantidad de riesgos desde una perspectiva única muy concreta, la de la privacidad y la protección de datos, construida a partir de un sistema que se centra fundamentalmente en la protección individual.

En primer lugar, existen diferentes situaciones en las que la propia protección de la privacidad es por sí misma insuficiente. Así, por ejemplo, los datos anonimizados no se incluyen en el marco de protección del RGPD y la Directiva 2016/680 y, sin embargo, es posible que en algunos casos se llegue a extraer información personal de datos supuestamente anonimizados. A mayor abundamiento, el uso de datos anonimizados también puede contribuir a la construcción de perfiles que, por ejemplo, perpetúen estereotipos perjudiciales para las personas pertenecientes a grupos desaventajados y que sí pueden ser aplicados en la toma de decisiones relativas a personas identificadas. Cabe también destacar que el marco jurídico en materia de protección de datos se centra, sobre todo, en los datos introducidos en los sistemas y no tanto en los resultados obtenidos. En este sentido, el RGPD no articula mecanismos de protección frente a las inferencias obtenidas del procesamiento de datos, sino que la protección se centra en la corrección y precisión de los datos introducidos inicialmente en el sistema y la forma en que estos son procesados.

En segundo lugar, los derechos individuales reconocidos a las personas interesadas y obligaciones de los responsables y encargados del tratamiento de datos que conforman el sistema de proceso tecnológico justo sitúan una serie de cargas sobre las personas cuyos datos son procesados que estas no siempre son capaces de asumir. En este sentido, incluso aunque las personas suelen expresar cierta preocupación y voluntad de proteger su privacidad y datos de carácter personal, sus actuaciones efectivas en relación a la misma no se corresponden en la mayor parte de los casos, a la postre, con dicha voluntad. Esto se debe, en parte, al fenómeno denominado “paradoja de la privacidad”, que refleja la forma en la que las personas no valoran de manera racional los daños futuros que pueden derivarse de compartir sus datos en el presente. Así, se da prioridad a los beneficios a corto plazo que generados por el acceso a los servicios digitales que, como contraprestación, exigen, solicitan o incluso toman sin informar de ello en exceso un generoso y creciente acceso a todos aquellos datos de carácter personal relacionados con la interacción o susceptibles de ser recopilados aprovechándola.

Asimismo, también cabe tener en cuenta las asimetrías informativas relativas a los datos que verdaderamente se comparten y al uso que se les da por las entidades que los recogen ya que las personas que comparten sus datos no son siempre conscientes de la información a la que están dando acceso. Esas asimetrías informativas se dan no solo con respecto a los datos que se comparten, sino que también se sitúan en el marco de los derechos que las personas tienen con respecto a sus propios datos. La amplia y compleja regulación en materia de protección de datos, en combinación con la baja percepción de los riesgos derivados de la recogida de datos personales, dificulta la concienciación social con respecto a la importancia de la privacidad y la comprensión de los derechos y posibilidades de protección de la esfera de la intimidad correspondiente a los datos personales.

En cualquier caso, cabe destacar que, en muchas ocasiones, las personas solo pueden elegir entre acceder a un servicio compartiendo sus datos o no acceder al mismo en absoluto y, teniendo en cuenta que el acceso a redes sociales y plataformas digitales se ha convertido en un elemento esencial de la vida social y laboral de cada vez más personas, esa supuesta libertad de elegir queda enormemente mermada.

En tercer lugar, el sistema de control general contenido en el RGPD y que pretende articularse como sistema de gobernanza general de los sistemas automatizados de procesamiento de datos, sencillamente, no es efectivo. En este sentido, cabe destacar que la elaboración y aprobación de códigos de conducta y mecanismos de certificación no requiere de la participación de terceros interesados (asociaciones o personas físicas), ya que el RGPD únicamente prevé la participación de asociaciones y otros organismos que representen a responsables y encargados del tratamiento de datos y de las autoridades relevantes. Además, cabe añadir que no se está realizando un seguimiento de la correcta implementación de las evaluaciones de impacto que deben ser llevadas a cabo de manera obligatoria cuando estos sistemas generen riesgos significativos para los derechos y libertades fundamentales.

Finalmente, la protección de los datos de carácter personal puede ser contraproducente para la protección de la igualdad y la no discriminación. Si se prohíbe la utilización de determinadas categorías de datos, como la religión o la raza, en el procesamiento de datos para la toma de decisiones, es mucho más complicado determinar si el algoritmo infiere las categorías protegidas de otros datos. Así, autorizar que los algoritmos empleen las categorías sospechosas en el procesamiento permite determinar hasta qué punto dichas categorías condicionan el resultado del algoritmo y qué otros datos se hallan vinculados con la

pertenencia a un grupo desaventajado. De lo contrario, aunque no resulte en un plano teórico totalmente imposible, sí es más complicado identificar si el algoritmo está infiriendo la pertenencia a un grupo desaventajado de otras categorías de datos. Asimismo, es más sencillo enseñar a los sistemas automatizados a no discriminar con base en las categorías sospechosas si dichos datos se introducen y consideran expresamente. De esta manera se puede enseñar al algoritmo a evitar confundir características (variables) de las que se infiere la pertenencia a un grupo desaventajado con variables que realmente tengan un valor predictivo en la determinación de la variable objetivo.

10. Una aproximación conjunta a los sistemas normativos en materia de igualdad y no discriminación y en materia de protección de datos puede aportar una solución provisional a las insuficiencias para la protección frente a la discriminación algorítmica que cada uno de dichos marcos jurídicos presenta cuando son considerados de manera individual

Uno de los problemas del marco jurídico antidiscriminatorio es el hecho de que la aplicación de las prohibiciones de discriminar solo puede llevarse a cabo después de que se haya producido la acción discriminatoria. Asimismo, como ya se ha indicado, la efectividad del marco jurídico antidiscriminatorio se puede ver menoscabada por la dificultad de detectar las situaciones de discriminación y de acceder a los órganos jurisdiccionales como consecuencia de la existencia de diversas trabas de índole burocrática, procesal y económica.

El marco europeo de protección de datos ofrece una serie de instrumentos que, en cierta medida, puede ayudar a completar las lagunas del marco antidiscriminatorio a la hora de lidiar con los riesgos para los derechos a la igualdad y a la no discriminación que generan los algoritmos. El RGPD y otras normas en materia de protección de datos ofrecen la posibilidad de llevar a cabo un control *ex ante* de los sistemas automatizados y refuerzan el control *ex post* que ofrecen los instrumentos jurídicos de lucha contra la discriminación. Los principios contenidos en el artículo 5 de la RGPD obligan a los responsables y encargados del tratamiento a garantizar que los datos de entrada no estén sesgados (exactitud) y que los datos se procesen "de forma lícita, justa y transparente", lo que introduce indirectamente la obligación de considerar, al menos, cómo la forma en que un sistema procesa los datos afecta a los derechos a la igualdad y a la no discriminación.

Asimismo, el principio de transparencia (también contenido en el art. 5 RGPD) cristaliza en los derechos al debido proceso y en los mecanismos de rendición de cuentas que figuran en el Reglamento. El sistema de derechos de los interesados reconocido en el Reglamento podría proporcionar los elementos necesarios para que las personas tomen conciencia de las decisiones automatizadas discriminatorias a las que podrían ser sometidas y demostrar algunos casos de discriminación algorítmica. Además, los mecanismos de rendición de cuentas y gobernanza, como los códigos de conducta, la certificación y las evaluaciones de impacto relativas a la protección de datos, no sólo podrían contribuir a lograr un control *ex ante* y continuo de los sistemas algorítmicos, sino que también ayudarían a subsanar las deficiencias de los sistemas de protección de los derechos fundamentales que principalmente imponen la carga al individuo.

También cabe destacar que considerando 71 del RGPD establece la obligación de los responsables del tratamiento de introducir las medidas necesarias para evitar efectos discriminatorios basados en categorías especiales de datos. A pesar de no incluirse dicho mandato en el articulado del Reglamento, dado que el art. 24.1 RGPD fija la obligación de establecer las medidas necesarias para garantizar el cumplimiento del Reglamento, teniendo en cuenta los posibles riesgos para los derechos y las libertades, cabe concluir que la noción de igualdad en el diseño se encuentra también indirectamente incluida en el marco normativo europeo en materia de protección de datos.

Incluso con sus muchas deficiencias y teniendo en cuenta las tensiones entre las normas en materia de protección de datos y las de protección de la igualdad y no discriminación, es innegable que la combinación de ambos marcos normativos ofrece un sistema más completo de protección contra la discriminación algorítmica. La combinación de ambos marcos jurídicos permite trazar los elementos estructurales de un modelo en el que basar las muy necesarias mejoras que el actual sistema normativo precisa para poder abordar la discriminación algorítmica de manera adecuada. Sin embargo, por el momento, y dado el evidente retraso de las disposiciones institucionales y jurídicas dirigidas a regular y controlar todos los riesgos que plantean los algoritmos y, en particular, la discriminación algorítmica, los operadores técnicos y jurídicos deberían centrarse en la combinación aplicada de los marcos de protección de datos, igualdad y lucha contra la discriminación a fin de lograr un sistema más eficaz que proteja los derechos a la igualdad y la no discriminación contra los peligros generados por la automatización de procesos de toma de decisiones.

11. Los riesgos y daños generados por los algoritmos justifican la necesidad de un sistema de regulación e intervención pública más desarrollado e intenso que el actualmente ofrecido por el marco jurídico en materia de protección de datos

El creciente uso de algoritmos genera daños a los derechos fundamentales a la protección de datos y a la igualdad y no discriminación, entre otros. Asimismo, la creciente automatización de procesos también produce daños a valores e intereses públicos propios de los Estados democráticos y de Derecho.

Cuando los algoritmos son utilizados por poderes públicos, es esencial, en primer lugar, reconocer los efectos jurídicos de estos sistemas sobre los destinatarios de los procedimientos total o parcialmente automatizados. En segundo lugar, y teniendo en cuenta el estado actual de las normas jurídicas específicas aplicables a los sistemas automatizados, es esencial que, por ahora, sean dotados de las garantías propias a los productos normativos tradicionales de predeterminación de la toma de decisiones públicas, esto es, las garantías propias de los reglamentos. Al fin y al cabo, los algoritmos contienen una serie de instrucciones predeterminadas que van aplicando al caso concreto. Es cierto que hay aspectos en los que difícilmente puede equipararse un algoritmo a un reglamento, sin embargo, calificarlos como tal, al menos de manera temporal, proporciona una serie de medidas de protección frente a los riesgos y daños que pueden generar los algoritmos y ofrece una base sobre la cual construir unas categorías jurídicas específicas aplicables a los sistemas automatizados.

12. El nuevo sistema de protección frente a los riesgos generados por los algoritmos debe incorporar una entidad independiente de supervisión, un sistema regulatorio basado en el riesgo y una serie de mecanismos adicionales que contribuyan a asegurar la regulación y control efectivos de los algoritmos

La propuesta realizada en la tesis parte, sobre todo, de tratar de garantizar en todo caso el objetivo último de la protección de la igualdad y no discriminación. Sin embargo, las estrategias propuestas a continuación aspirar también a sentar las bases para crear, además, un sistema efectivo de control y regulación de los diferentes riesgos generados por los algoritmos, especialmente cuando son empleados en el ámbito de la toma de decisiones por parte de los poderes públicos.

12.1. Autoridad única de control algorítmico

El marco jurídico e institucional en materia de protección de datos resulta insuficiente para abordar todos los riesgos surgidos del creciente uso de sistemas algorítmicos. Sin embargo, la red institucional creada en la Unión Europea con el objetivo de proteger los datos personales de la ciudadanía puede servir como base para que, a partir de estas agencias independientes o por medio de nuevas estructuras institucionales especiales, se asuman paulatinamente mayores competencias respecto del control de los algoritmos. Sobre todo, teniendo en cuenta los importantes riesgos que estos sistemas generan para los derechos a la igualdad y a la no discriminación, es importante que la composición de estas instituciones cuente con personal especializado en estudios de igualdad así como con especialistas en el campo de la ética algorítmica.

12.2. Regulación y control de los algoritmos mediante un sistema basado en el riesgo

El sistema de control de los algoritmos basado en el riesgo establecería cinco niveles de riesgo.

- Los sistemas que produzcan unos riesgos inasumibles, como las armas autónomas (que no automáticas), deberían quedar directamente y en todo caso prohibidos.

- Los sistemas que generen riesgos elevados (algoritmos utilizados en la contratación de personal, por ejemplo), en el caso que nos ocupa, para los derechos a la igualdad y la no discriminación, deberán contar con una autorización previa de acceso al mercado y posteriores auditorías públicas periódicas. Los algoritmos de esta clase empleados por las Administraciones públicas deberán, además, cumplir con una serie de requisitos adicionales específicos, incluyendo un mayor nivel de transparencia, debido a su capacidad de preconfigurar normativamente el ejercicio del poder público y su alta capacidad de afección a derechos e intereses de los ciudadanos. Así, deberían establecerse registros públicos de algoritmos en los que apareciesen públicados los algoritmos empleados por el sector público, como ya hace la ciudad de Ámsterdam.

- Los sistemas que generen riesgos medios (del estilo de los algoritmos utilizados en la adaptación de precios que usan habitualmente las grandes plataformas de comercio electrónico para optimizar el cruce de oferta y demanda, así como sus beneficios) deberán, de manera obligatoria, ser certificados por entidades reconocidas por las Autoridades de control algorítmico. Este proceso de certificación deberá asegurar,

entre otros extremos, que los algoritmos no vulneran en ningún caso los derechos a la igualdad y a la no discriminación y que aseguran un trato equitativo a los grupos desfavorecidos. Desde un primer momento se deberá entregar a las referidas Autoridades toda la documentación relativa al *software* empleado, que se conservará por esta. Se realizarán procesos de recertificación posteriores con carácter periódico.

- Los sistemas que generen riesgos de nivel bajo (por ejemplo, algoritmos utilizados en sistemas GPS) también deberán obtener una certificación y someterse a procesos de re-certificación periódicos.

- Los sistemas con riesgo prácticamente inexistente (algoritmos empleados en máquinas expendedoras y, en general, programas que realicen operaciones muy sencillas y mecánicas) no deberán superar más controles que aquellos a los que se deban someter los productos en los que se incorporan.

12.3. Establecimiento de un sistema de "mejores técnicas disponibles"

El desarrollo de sistemas algorítmicos cada vez más complejos que contribuyen a perpetuar determinadas estructuras de desigualdad social ha venido también acompañado de una incipiente rama académica y profesional dedicada específicamente a la detección de la discriminación algorítmica y a la creación de algoritmos que se pretenden no discriminatorios. Es por ello que el establecimiento de un sistema europeo de mejores técnicas disponibles en la materia permitiría un conocimiento homogéneo y compartido de las posibilidades respecto de la protección de la igualdad y la garantía de la no discriminación, así como de otros riesgos generados por los algoritmos.

También cabe considerar que, a pesar del imparable desarrollo tecnológico en el ámbito de las técnicas de recogida y procesamiento de datos, en ocasiones todavía resulta muy costoso acudir a las mejores técnicas que logren los resultados que sean además lo suficientemente eficientes, con la mayor transparencia, el mayor posible respeto a los derechos a la igualdad y a la no discriminación y a otros derechos fundamentales, etc. Es por ello que el establecimiento de un sistema europeo de "mejores técnicas disponibles" permitiría dar a conocer cuáles deben ser los algoritmos que se pueden emplear en según qué contextos, facilitando también la tarea de los tribunales en el enjuiciamiento de supuestos de conflictos entre derechos e intereses en riesgo por la toma de decisiones automatizadas.

12.4. La contratación pública como mecanismo para prevenir los riesgos en el uso de algoritmos por los sectores público y privado

Muchos de los sistemas automatizados utilizados por las Administraciones públicas proceden, en realidad, del sector privado. El sector público deberá aprovechar los mecanismos disponibles en la normativa de contratos del sector público para imponer determinadas condiciones relativas tanto a la composición y organización de las empresas licitadoras como concernientes a las características de los propios algoritmos. De esta manera se fomentará, de manera indirecta, que los algoritmos utilizados también en el sector privado se elaboren empleando exigencias y parámetros parecidos. Dada la ya larga tradición europea estableciendo unos mínimos comunes en materia de contratación pública, esta labor no supone además un gran cambio ni expansión de las materias de las que tradicionalmente se ha considerado competente la Unión Europea.

12.5. Detección de la discriminación y la desigualdad mediante algoritmos

Los sistemas algorítmicos pueden ayudar a la producción de importantes avances en la promoción y protección de la igualdad y en la lucha contra la discriminación. La capacidad que estos sistemas tienen para analizar enormes cantidades de datos y para generar resultados altamente precisos sobre la realidad debería ser empleada para detectar aquellas situaciones de vulnerabilidad y desventaja de determinados grupos poblacionales e incluso como una herramienta particularmente útil para identificar y atajar la discriminación estructural y las situaciones específicas de discriminación interseccional.

Asimismo, debe detenerse la actual tendencia a la automatización de sistemas de ayudas públicas y servicios sociales que, en lugar de ofrecer una mayor y más efectiva protección a las personas vulnerables, desvirtúan en no pocos casos los objetivos de las herramientas y políticas propias del Estado de bienestar, convirtiendo estos sistemas en meros mecanismos de detección de fraude.

12.6. Empoderamiento de las personas en la gestión de sus datos

Actualmente, no existen suficientes incentivos para que los individuos se preocupen por salvaguardar su privacidad todo lo posible. Una persona que no percibe un peligro real en compartir sus datos no va a preocuparse por identificar y controlar con qué proveedores se

comparten sus datos cada vez que entra en una página web y deseleccionarlos todos, aunque formalmente se les dé la oportunidad de hacerlo. Incluso cuando la protección de los datos es procedimentalmente más sencilla, resulta costoso prestar atención a este elemento cuando lo que se pretende es acceder a un servicio con relativa rapidez. Asimismo, y por descontado, una persona no va a leerse los términos y condiciones de uso o la política de privacidad de cada empresa cuya página web visita o cuyo producto digital adquiere. Debe, por tanto, trabajarse en fomentar arquitecturas del diseño que se centren en la protección de los intereses de los individuos por defecto y sin que sea necesaria una intervención activa de éstos, y menos todavía una intervención que requiere de un nivel de conocimiento y dedicación elevados.

12.7. Comunicación y cooperación entre disciplinas

Quizás uno de los elementos regulatorios de futuro necesarios, absolutamente esenciales para un sistema efectivo de control y regulación de los sistemas algorítmicos, es el fomento de la comunicación y colaboración entre disciplinas. Las diferencias entre la forma en que se preparan, desarrollan y aprueban las normas jurídicas ordinarias y la forma en la que funcionan los algorítmicos y los programas de *software* producen rupturas significativas que impiden y escapan en la forma tradicional de abordar los problemas sociales por el Derecho. Asimismo, las personas dedicadas a la programación de sistemas están acostumbradas a unas formas de trabajo y de desarrollo de productos que difícilmente encajan con la consecución del respeto a los derechos fundamentales e intereses públicos y el control de la conformidad a Derecho de los algoritmos como fines prioritarios. Es por ello que la creciente incidencia de los sistemas automatizados en la vida política, social y económica requiere como *prius* inexcusable, para su correcta regulación y utilización, de la comunicación y cooperación entre personas expertas en una amplia gama de disciplinas y, sobre todo, entre aquellos encargados de sus respectivos desarrollos jurídicos y tecnológicos.

BIBLIOGRAPHY AND SOURCES

1. Articles

ALBERTÍN, P., CUBELLS, J. & PEÑARANDA, M. C., "A feminist law meets an androcentric criminal justice system: gender-based violence in Spain", *Feminist Criminology*, 2018, pp. 1-27.

ALEXY, R. O., "On constitutional rights to protection", *Legisprudence*, vol. 3, No. 1, 2009, pp. 1-17.

- "Constitutional rights and proportionality", *Journal for Constitutional Theory and Philosophy of Law*, vol. 22, 2014, pp. 51-65.

ALIMADADI, A., ARYAL, A., MANANDHAR, I., MUNROE, P. B., JOE, B. & CHENG, X., "Artificial intelligence and machine learning to fight COVID-19", *Physiological Genomics*, vol. 52, 2020, pp. 200-202.

ALKSNIS, C., DESMARAIS, S. & CURTIS, J., "Workforce segregation and the gender wage gap: is 'women's' work valued as highly as 'men's'?", *Journal of Applied Social Psychology*, vol. 38, No. 6, 2008, pp. 1416-1441.

ALLEN, R. & MASTERS, D., "Artificial Intelligence: the right to protection from discrimination caused by algorithms, machine learning and automated decision-making", *ERA Forum*, vol. 2020, No. 4, 2020, pp. 586-598.

ALLHUTTER, D., CECH, F., FISCHER, F., GRILL, G. & MAGER, A. "Algorithmic profiling of job seekers in Austria: how austerity politics are made effective", *Frontiers in Big Data*, vol. 3, 2020, pp. 1-17.

ÁLVAREZ GARCÍA, V., "Introducción a los problemas jurídicos de la normalización industrial: normalización industrial y sistema de fuentes (1)", *Revista de Administración Pública*, No. 147, 1998, pp. 307-336.

ANNANY, M. & CRAWFORD, K., "Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability", *New Media & Society*, vol. 20, No. 3, 2018, pp. 973-989.

AÑÓN ROIG, M. J., "Principio antidiscriminatorio y determinación de la desventaja", *Isonomía: Revista de Teoría y Filosofía del Derecho*, No. 39, 2013b, pp. 127-157.

APPEL, M., WEBBER, S. & KRONBERGER, N., "The influence of stereotype threat on immigrants: review and meta-analysis", *Frontiers in Psychology*, 2015, pp. 1-15.

AVERY, R. B., BOSTIC, R. W., CALEM, P. S., & CANNER, G. B., "Credit scoring: statistical issues and evidence from credit-bureau files", *Real Estate Economics*, vol. 28, 2000, pp. 523-547.

ARZOZ SANTIESTEBAN, X., "La eficacia del CEDH en las relaciones entre particulares", *Anuario de la Facultad de Derecho de la Universidad Autónoma de Madrid*, No. 21, 2017, pp. 149-174.

ATHEY, S., CATALINI, C. & TUCKER, C., "The digital privacy paradox: small money, small costs, small talk", *MIT Sloan Research Paper No. 5196-17*, 2017. Available on 21st May 2019 at: https://papers.ssrn.com/

BALAGUER CALLEJÓN, M. L., "Igualdad y discriminación sexual en la jurisprudencia del TC", *Revista de Derecho Político*, No. 33, 1991, pp. 99-124.

BALKIN, J. M. & SIEGEL, R. B., "The American civil rights tradition: anticlassification or antisubordination?", *University of Miami Law Review*, vol. 58, No. 1, 2003, pp. 9-34.

BAMBAUER, J. & ZARSKY, T., "The algorithm game", *Notre Dame Law Review*, vol. 94, No. 1, 2018, pp. 1-48.

BAMBERGER, K. A., "Regulation as delegation: private firms, decisionmaking, and accountability in the administrative state", *Duke Law Journal*, vol. 56, No. 2, 2006, pp. 377-468.

BAR-GILL, O., "Algorithmic price discrimination when demand is a function of both preferences and (mis)perceptions", *The University of Chicago Law Review*, vol. 86, No. 2, 2019, pp. 217-254.

BARNARD, C. & HEPPLE, B., "Indirect discrimination: interpreting Seymour-Smith", *Cambridge Law Journal*, vol. 58, No. 2, 1999, pp. 399-412.

BAROCAS, S. & SELBST, A. D., "Big data's disparate impact", *California Law Review*, vol. 104, No. 3, 2016, pp. 671-732.

- "The intuitive appeal of explainable machines", *Fordham Law Review*, vol. 87, No. 3, 2018, pp. 1085-1139.

BARON, J. & SPULBER, D. F., "Technology standards and standard setting organizations: introduction to the Searle Center database", *Journal of Economics & Management Strategy*, No. 27, 2018, pp. 462-503.

BARRÈRE UNZUETA, M. A., "Problemas del derecho antidiscriminatorio: subordinación versus discriminación y acción positiva versus igualdad de oportunidades", *Cuadernos Electrónicos de Filosofía del Derecho*, No. 9, 2003a, pp. 1-26.

- "Igualdad y 'discriminación positiva': un esbozo de análisis teórico-conceptual", *Cuadernos Electrónicos de Filosofía del Derecho*, No. 9, 2003b, pp. 1-27.

BARRÈRE UNZUETA, M. A. & MORONDO TARAMUNDI, D., "Subordiscriminación y discriminación interseccional: elementos para una teoría del derecho antidiscriminatorio", *Revista de Filosofía Jurídica y Política*, vol. 45, 2011, pp. 15-42.

BASHI, V., "Racial categories matter because racial hierarchies matter: a commentary", *Ethnic and Racial Studies*, vol. 21, No. 5, 1998, pp. 959-968.

BAYAMLIOĞLU, E., "Transparency of automated decisions in the GDPR: an attempt for systemisation", 2018. Available on 9th May 2019 at: https://ssrn.com/

BAYÓN, M. C., "La construcción del otro y el discurso de la pobreza: narrativas y experiencias desde la periferia de la ciudad de México", *Revista Mexicana de Ciencias Políticas y Sociales*, vol. 60, No. 223, 2015, pp. 357-376.

BELLOVIN, S. M., HUTCHINS, R. M., JEBARA, T., ZIMMECK, S., "When enough is enough: location tracking, mosaic theory, and machine learning", *NYU Journal of Law & Liberty*, vol. 8, 2014, pp. 555-628.

BENBOUZID, B., "Des crimes et des séismes: la police prédictive entre science, technique et divination", *La Découverte*, No. 206, 2017, pp. 95-123.

BEN-SHAHAR, O., "Data pollution", *Journal of Legal Analysis*, vol. 11, 2019, pp. 104-159.

BENNET MOSES, L. & CHAN, J., "Algorithmic prediction in policing: assumptions, evaluation, and accountability", *Policing and Society*, vol. 28, No. 7, 2018, pp. 806-822.

BENT, J. R., "Is algorithmic affirmative action legal?", *The Georgetown Law Journal*, vol. 108, 2020, pp. 803-853.

BERK, R., HEIDARI, H., JABBARI, S., KEARNS, M. & ROTH, A., "Fairness in criminal justice risk assessments: the state of the art", *Sociological Methods and Research*, 2018, pp. 1-24.

BERNS, S., "Liberalism and the privatised family: the legacy of Rousseau", *Res Publica*, vol. 11. No. 2, 2005, pp. 125-155.

BERRIDGE, C. W., & MARTINSON, M., "Valuing old age without leveraging ableism", *Generations,* vol. 41, No. 4, 2018, pp. 83-91.

BILBAO UBILLOS, J. M., "La consolidación dogmática y jurisprudencial de la *Drittwirkung*: una visión de conjunto", *Anuario de la Facultad de Derecho de la Unviersidad Autónoma de Madrid*, No. 21, 2017, pp. 43-74.

BILGE, S., "Saving intersectionality from feminist intersectionality studies", *Du Bois Review*, vol.10, No. 2, 2013, pp. 405-424.

BINNS, R., "Data protection impact assessments: a meta-regulatory approach", *International Data Privacy Law*, vol. 7, No. 1, 2017, pp. 22-35.

BLACK, J., "The emergence of risk-based regulation and the new public risk management in the United Kingdom", *Public Law*, No. 3 (Autumn), 2005, pp. 512-548.

- "Learning from regulatory disasters", *LSE Law, Society and Economy Working Papers* 24/2014, 2014, pp. 1-18. Available on 6th March 2020 at: http://eprints.lse.ac.uk/

BOHREN, J. A., HAGGAG, K., IMAS, A. & POPE, D. G., "Inaccurate statistical discrimination", *NBER* Working Paper No. 25935, 2019. Available on 3rd July 2019 at: https://www.nber.org/

BOIX PALOP, A., "El equilibro entre los derechos del artículo 18 de la Constitución, el «derecho al olvido» y las libertades informativas tras la Sentencia Google", *Revista General de Derecho Administrativo*, No. 38, 2015.

- "Los algoritmos son reglamentos: la necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones", *Revista de Teoría y Método*, vol. 1, 2020, pp. 223-270.

BORNSTEIN, S., "Antidiscriminatory algorithms", *Alabama Law Review*, vol. 70, No. 2, 2019, pp. 519-572.

BOVENBERG, J., PELOQUIN, D., BIERER, B., BARNES, M. & KNOPPERS, B. M., "How to fix the GDPR's frustration of global biomedical research", *Science*, vol. 370, No. 6512, pp. 40-42.

BRAUNEIS, R. & GOODMAN, E. P., "Algorithmic transparency for the smart city", *Yale Journal of Law & Technology*, vol. 20, 2018, pp. 103-176.

BRENNAN-MARQUEZ, K., "'Plausible cause': explanatory standards in the age of powerful machines"**,** *Vanderbilt Law Review*, vol. 17, No. 4, 2017, pp. 1249-1301.

BRKAN, M., "Do algorithms rule the world? Algorithmic decision-making in the framework of the GDPR and beyond", 2017, pp. 1-29. Available on 9th May 2019 at: https://ssrn.com/

- "Do algorithms rule the world? Algorithmic decision-making in the framework of the GDPR and beyond", *International Journal of Law and Information Technology*, vol. 27, No. 2, 2019 pp. 91-121.

BURRELL, J., "How the machine 'thinks': understanding opacity in machine learning algorithms", *Big Data & Society*, vol. 3, No. 1, 2016, pp. 1-12.

BURT, A. & VOLCHENBOUM, S., "How healthcare changes when algorithms start making diagnoses", *Harvard Business Review*, 8th March 2018. Available on 25th March 2019 at: https://hbr.org/

BYFIELD, N. P., "Race science and surveillance: police as the new race scientists", *Social identities: journal for the study of race, nation and culture*, vol. 25, No. 1, 2018, pp. 91-106.

BYGRAVE, L. A., "Minding the machine: Article 15 of the EC data protection directive and automated profiling", *Computer Law & Security Review*, vol. 17, No. 1, 2001, pp. 17-24.

- "EU data protection law falls short as desirable model for algorithmic regulation", *Centre for Analysis Risk and Regulation Discussion Papers*, No. 85, 2017, pp. 31-33.

CABALLÉ-PÉREZ, M., VILLALBA-GARCÍA, D., SANTOS-HERMOSO, J., LÓPEZ-OSSORIO, J. J. & GONZÁLEZ-ÁLVAREZ, J. L., "El quebrantamiento de las órdenes de protección en violencia de género: análisis de los indicadores de riesgo mediante el formulario vpr4.0", *Anuario de Psicología Jurídica*, No. 30, 2020, pp. 63-72.

CALDERÓN CARRERO, J. M., "El encuadramiento legal y límites del uso de herramientas de inteligencia artificial con fines de control fiscal: Análisis de la decisión del Consejo Constitucional francés de 27 de diciembre de 2019 (Décision n.º 2019-796 DC), sobre la Ley de Presupuestos 2020", *CEF Revista de Contabilidad y Tributación*, Nº. 444, 2020, pp. 119-128.

CAMPBELL, F. A. K., "Exploring internalized ableism using critical race theory", *Disability & Society*, vol. 23, No. 2, 2008, pp. 151-162.

CAPDEFERRO VILLAGRASA, O., "El análisis de riesgos como mecanismo central de un sistema efectivo de prevención de la corrupción. En particular, el sistema de alertas para la prevención de la corrupción basado en inteligencia artificial", *Revista Internacional de Transparencia e Integridad*, No. 6, 2018, pp. 1-7.

CARACCIOLO DI TORELLA, E., "The principle of gender equality, the goods and services directive and insurance: A conceptual analysis", *Maastricht Journal of European and Comparative Law*, vol. 13, No. 3, pp. 339-350.

CARUANA, M. M., "The reform of the EU data protection framework in the context of the police and criminal justice sector: harmonisation, scope, oversight and enforcement", *International Review of Law, Computers & Technology*, 2017, pp. 1-22.

CASTRO, D. & MCLAUGHLIN, M., "Ten ways the precautionary principle undermines progress in artificial intelligence", *Information Technology & Innovation Foundation*, 4th February 2019. Available on 14th May 2020 at: https://itif.org/

CATH, C., WACHTER, S., MITTELSTADT, B., TADDEO, M. & FLORIDI, L., "Artificial intelligence and the 'good society': the US, EU and UK approach", *Science and Engineering Ethics*, vol. 24, No. 2, 2018, pp. 505-528.

CEA D'ANCONA, M. A., "Immigration as a threat: explaining the changing pattern of xenophobia in Spain", *International Immigration & Integration*, vol. 17, 2016, pp. 569-591.

CECCHI-DIMEGLIO, P., "How gender bias corrupts performance reviews, and what to do about it", *Harvard Business Review*, 12th April 2017. Available on 7th April 2019 at: https://hbr.org/

CELIOUS, A. & OYSERMAN, D., "Race from the inside: an emerging heterogeneous race model", *Journal of Social Issues*, Vol. 57, No. 1, 2001, pp. 149-165.

CHAR, D. S., SHAH, N. H. & MAGNUS, D., "Implementing machine learning in health care – addressing ethical challenges", *The New England Journal of Medicine*, vol. 378, No.11, 2018, pp. 981-983.

CHEN, M., HAO, Y., HWANG, K., WANG, L. & WANG, L., "Disease prediction by machine learning over big data from healthcare communities", *IEEE Access*, vol. 5, 2017, pp. 8869-8879.

CHOULDECHOVA, A., "Fair prediction with disparate impact: a study of bias in recidivism prediction instruments", 2016. Available on 9th April 2019 at: https://arxiv.org/

CITRON, D. K., "Technological due process", *Washington University Law Review*, vol. 85, No. 6, 2008, pp. 1249-1313.

CITRON, D. K., & PASQUALE, F, "The scored society: due process for automated predictions", *Washington Law Review Online*, vol. 89, 2014, pp. 1-33.

CLIFFORD, S., "The capacity contract: Locke, disability, and the political exclusion of "Idiots", *Politics, Groups and Identities*, vol. 2, No. 1, 2014, pp. 90-103.

COBBE, J., "Administrative law and the machines of government: judicial review of automated public-sector decision-making", *Legal Studies*, vol. 39, No. 4, 2019, pp. 363-655.

COGLIANESE, C., "Optimizing Regulation for an optimizing economy", *University of Pennsylvania Journal of Law & Public Affairs*, vol. 4, No. 1, 2018, pp. 11-13.

COGLIANESE, C. & LEHR, D., "Regulating by robot: administrative decision making in the machine-learning era", *The Georgetown Law Journal*, vol. 105, No. 5, 2017, pp. 1147-1223.

COOK, C., DIAMOND, R., HALL, J., LIST, J. A. & OYER, P., "The gender earnings gap in the gig economy: evidence from over a million rideshare drivers", *NBER* Working Paper No. w24732, 2019, pp. 1-62. Available on 2nd March 2020 at: https://www.nber.org/

CORBETT-DAVIES, S. & GOEL, S., "The measure and mismeasure of fairness: a critical review of fair machine learning", 2018, pp. 1-25. Available on 6th May 2020 at: https://arxiv.org/

COTINO HUESO, L., "Riesgos e impactos del big data, la inteligencia artificial y la robótica: enfoques, modelos y principios de la respuesta del derecho", *Revista General de Derecho Administrativo*, No. 50, 2019.

CRAWFORD, K., "Think again: big data", *Foreign Policy*, 10th May 2013. Available on 25th September 2018 at: https://foreignpolicy.com/

CRAWFORD, K. & SCHULTZ, J., "Big data and due process: toward a framework to redress predictive privacy harms", *Boston College Law Review*, vol. 55, No. 1, 2014, pp. 93-128.

CRENSHAW, K., "Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics", *University of Chicago Legal Forum*, No. 1, 1989, pp. 139-167.

CROSBY, F. J., IYER, A. & SINCHARO, S., "Understanding affirmative action", *Annual Review of Psychology*, vol. 57, 2006, pp. 585-611.

CUKIER, K., & MAYER-SCHOENBERGER, V., "The rise of big data: How it's changing the way we think about the world", *Foreign Affairs*, vol. 92, No. 93, 2013, pp. 28-40.

CUYPERS, D., "Religion, discrimination, the head scarf and labour law", *ERA Forum*, vol. 19, 2019, pp. 415-448.

DATTA, A., FREDRIKSON, M., KO, G., MARDZIEL, P. & SEN, S., "Proxy non-discrimination in data-driven systems: theory and experiments with machine learnt programs", 2017, pp. 1-14. Available on 15th February 2019 at: https://arxiv.org/

DE BURCA, G., "The trajectories of European and American antidiscrimination law", *The American Journal of Comparative Law*, vol. 60, No. 1, 2012, pp. 1-22.

DE HERT, P., "Data protection as bundles of principles, general rights, concrete subjective rights and rules: piercing the veil of stability surrounding the principles of data protection", *European Data Protection Law Review*, vol. 3, No. 2, 2017, pp. 160-179.

DE LA CUEVA, J., "Código fuente, algoritmos y fuentes del derecho", *El Notario del Siglo XXI*, No. 77, 2018. Available on 18th June 2020 at: https://www.elnotario.es/

- "El derecho a no ser gobernados mediante algoritmos secretos", *El Notario del Siglo XXI*, No. 87, 2019. Available on 18th February 2020 at: https://www.elnotario.es/

DE LA SIERRA MORÓN, S., "Inteligencia artificial y justicia administrativa: una aproximación desde la teoría del control de la Administración pública", *Revista General de Derecho Administrativo*, No. 53, 2020, pp. 1-19.

DILL, K. E. & THILL, K. P., "Video game characters and the socialization of gender roles: young people's perceptions mirror sexist media depictions", *Sex Roles: A Journal of Research*, vol. 57, No. 11-12, 2007, pp. 851-864.

DIAKOPOULOS, N., "Accountability in algorithmic decision making", *Communications of the ACM*, 2016, vol. 59, No. 2, 2016, pp. 56-62.

DOLEAC, J. L. & HANSEN, B., "The unintended consequences of 'ban the box': statistical discrimination and employment outcomes when criminal histories are hidden", *Journal of Labour Economics*, vol. 38, No. 2, 2020, pp. 321-374.

DOYAL, L., "Sex, gender, and health: the need for a new approach", *British Medical Journal*, vol. 323, 2001, pp. 1061-1063.

DRECHSLER, L. & BENITO SÁNCHEZ, J. C., "The price is (not) right: data protection and discrimination in the age of pricing algorithms", *European Journal of Law and Technology*, vol. 9, No. 3, 2018, pp. 1-23.

DRESSEL, J. & FARID, H., "The accuracy, fairness, and limits of predicting recidivism", *Science Advances*, vol. 4, No. 1, 2018.

DREYER, S. & SCHULZ, W., "The General Data Protection Regulation and automated decision-making: will it deliver?", *Bertelsmann Stiftung*, 2019. Available on 7th May 2019 at: https://www.bertelsmann-stiftung.de/

EAGLIN, J. M., "Constructing recidivism risk", *Emory Law Journal*, vol. 67, No. 1, 2017, pp. 59-122.

EDWARDS, H. & STORKEY, A., "Censoring representations with an adversary", 2015, pp. 1-14. Available on 8th February 2020 at: https://arxiv.org/

EDWARDS, L. & VEALE, M., "Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for", *Duke Law & Technology Review*, vol. 16, 2017, pp. 18-84.

- "Enslaving the algorithm: from a 'right to an explanation' to a 'right to better decisions'?", *IEEE Security & Privacy*, vol. 16, No. 3, 2018, pp. 46-54.

EL EMAM, K. & ÁLVAREZ, C., "A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques", *International Data Privacy Law Journal*, vol. 5, No. 1, 2015, pp. 73-87.

ELEGIDO, J. M., "The ethics of price discrimination", *Business Ethics Quarterly*, vol. 21, No. 4, 2011, pp. 633-660.

EPSTEIN, R. & ROBERTSON, R. E., "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections", *PNAS*, vol. 112, No. 33, 2015, pp. E4512-E4521.

EREL, I., STERN, L. H., TAN, C. & WEISBACH, M. S., "Selecting directors using machine learning" *Fisher College of Business Working Paper No. 2018-03-005; European Corporate Governance Institute (ECGI) - Finance Working Paper No. 605/2019*, 2019. Available on 7th July 2019 at: https://ssrn.com/

ESKENS, S. E., "Profiling the European consumer in the internet of things: how will the General Data Protection Regulation apply to this form of personal data processing, and how should it?", 22nd March 2016. Available on 19th March 2020 at: https://papers.ssrn.com/

ESQUIVEL, V., "Efficiency and gender equality in growth theory: simply add-ons?", *Canadian Journal of Development Studies / Revue canadienne d'études du développement*, vol. 38, No. 4, 2017, 547-552.

FALIAGKA, E., ILIADIS, L.,KARYDIS, I., RIGOU, M., SIOUTAS, S., TSAKALIDIS, A. & TZIMAS, G., "On-line consistent ranking on e-recruitment: seeking the truth behind a well-formed CV", *Artificial Intelligence Review*, No. 42, 2014, pp. 515-528.

FAYYAD, U. M., PIATESKY-SHAPIRO, G., & SMYTH, P., "From data mining to knowledge discovery in databases", *AI Magazine*, vol. 17, No. 3, 1996, pp. 37-54.

FERGUSON, A. G., "Big data and predictive reasonable suspicion", *University of Pennsylvania Law Review*, vol. 163, No. 2, 2015, pp. 327-410.

- "Predictive prosecution", *Wake Forest Law Review*, vol. 51, No. 3, 2016, pp. 705-744.

FERNÁNDEZ LLERA, R. & MUÑIZ PÉREZ, M., "Colegios concertados y selección de escuela en España: un círculo vicioso", *Presupuesto y Gasto Público*, vol. 67, 2012, pp. 97-118.

FISHER LAVELL, E., "Beyond charity: social class and classism in counselling", *Canadian Journal of Counselling and Psychotherapy*, vol. 48, No. 3, 2014, pp. 231-250.

FLORIDI, L., "Mature information societies – a matter of expectations", *Philosophy and Technology*, vol. 29, No. 1, 2016a, pp. 1-4.

- "On human dignity as a foundation for the right to privacy", *Philosophy and Technology*, vol. 29, 2016b, pp. 307-312.

FOSTER, S. L., "Rawls, race and reason", *Fordham Law Review*, vol. 72, No. 5, 2004, pp. 1715-19.

FRIEDLER, S. A. *et al*, "On the (im)possibility of fairness", 2016. Available on 15th June 2020 at: https://arxiv.org/

FRIEDMAN, B. & NISSEMBAUM, H., "Bias in computer systems", *ACM Transactions on Information Systems*, vol. 14, No. 3, 1996, pp. 330-347.

GARRIDO FALLA, F., "El concepto de servicio público en el derecho español", *Revista de Administración Pública*, No. 135, 1994, pp. 7-36.

GELLERT, R., "Understanding the notion of risk in the General Data Protection Regulation", *Computer Law & Security Review***,** vol. 34, No. 2, 2018, pp. 279-288.

GERARDS, J., "The discrimination grounds of article 14 of the European Convention on Human Rights", *Human Rights Law Review*, vol. 13, No. 1, 2013, pp. 99-124.

GERARDS, J. & GLAS, L. R., "Access to justice in the European Convention of Human Rights system", *Netherlands Quarterly of Human Rights*, vol. 35, No. 1, 2017, pp. 11-30.

GIANFRANCESCO, M. A., TAMANG, S., YAZDANY, J. & SCHMAJUK, G., "Potential biases in machine learning algorithms using electronic health record data", *JAMA Internal Medicine*, vol. 178, No. 11.

GIEBLER, H. & MERKEL, W., "Freedom and equality in democracies: is there a trade-off?", *International Political Science Review*, vol. 37, No. 5, 2016, pp. 594-605.

GILLIS, T. B. & SPIESS, J. L., "Big data and discrimination", *The University of Chicago Law Review*, vol. 86, No. 2, 2019, pp. 459-487.

GOLDIN, C. & ROUSE, C., "Orchestrating impartiality: the impact of "blind" auditions on female musicians", *The American Economic Review*, vol. 90, No. 4, 2000, pp. 715-741.

GONÇALVES, M. E., "The risk-based approach under the new EU data protection regulation: a critical perspective", *Journal of Risk Research*, vol. 23, No. 2, 2020, pp. 139-152.

GONZÁLEZ ALVÁREZ, J. L., "Sistema de seguimiento integral en los casos de violencia de género (sistema viogén)", *Cuadernos de la Guardia Civil: Revista de Seguridad Pública*, No. 56, 2018, pp. 83-102.

GONZÁLEZ RAMOS, A. M. & TORRADO MARTÍN-PALOMINO, E., "Objectification and marketisation of women: technologies as instrument of violence", *Sociología y Tecnociencia*, vol. 9. No 1, 2019, pp. 1-8.

GOODMAN, B. & FLAXMAN, S., "European Union regulations on algorithmic decision making and a "right to explanation", *AI Magazine*, vol. 38, No 3, 2017, pp. 50-57.

GREENE, J., "Zero tolerance: a case study of police policies and practices in New York City", *Crime and Delinquency*, vol. 45, No. 2, 1999, pp. 171-187.

GRIMMELMANN, J. y WESTREICH, D., "Incomprehensible discrimination", *California Law Review Online*, vol. 7, 2017, pp. 164-177.

GUZELLA, T. S. & CAMINHAS, W. M., "A review of machine learning approaches to spam filtering", *Expert Systems with Applications*, vol. 36, No. 7, pp. 10206-10222.

HACKER, P., "Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law", *Common Market Law Review*, vol. 55, No. 4, 2018, pp. 1143-1186.

HADFIELD, G. K., "Feminism, fairness, and welfare: an invitation to feminist law and economics", *Annual review of law and social science*, vol. 1, 2005, pp. 285-306.

HAND, D. J., "Data mining: statistics and more?", *The American Statistician*, vol. 52, No. 2, 1998, pp. 112-118.

- "Classifier Technology and the illusion of Progress", *Statistical Science,* vol. 21, No. 1, 2006, pp. 1-15.

HARAWAY, D., "Situated knowledges: the science question in feminism and the privilege of partial perspective", *Feminist Studies,* vol. 14, No. 3 pp. 575-599.

HARNISH, A., "Ableism and the Trump phenomenon", *Disability & Society*, vol. 32, No.3, 2017, pp. 423-428.

HARTZOG, W., "The inadequate, invaluable fair information practices", *Maryland Law Review*, vol. 76, No. 4, 2017, pp. 952-983.

HENDERSON, L., HERRING, C., HORTON, H. D. & THOMAS, M., "Credit where credit is due?: race, gender, and discrimination in the credit scores of business startups" *The Review of Black Political Economy*, vol. 42, pp. 459-479.

HILDEBRANDT, M. & KOOPS, B. J., "The challenges of ambient law and legal protection in the profiling era", *The Modern Law Review*, vol. 73, No. 3, 2010, pp. 428-460.

HIRSCH, D. D., "The law and policy of online privacy: Regulation, self-regulation, or co-regulation", Seattle University Law Review, vol. 34, No. 2, 2011, pp. 439-480.

HO, A., "The individualist model of autonomy and the challenge of disability", *Journal of Bioethical Enquiry*, vol. 5, No. 2-3, 2008, pp. 193-207.

HODGE, N., "Lives worthy of life: the everyday resistance of disabled people", *Journal of Applied Hermeneutics*, 2016, pp. 1-7.

HOLZER, H. J., RAPHAEL, S., & STOLL, M. A., "Perceived criminality, criminal background checks, and the racial hiring practices of employers, *Journal of Law and Economics*, vol. 49, No. 2, 2006, pp. 451-480.

HOPKINS, P., "Social geography II: islamophobia, transphobia, and sizism", *Progress in Human Geography*, 2019, pp. 1-12.

HORDERN, V. "How do you solve a problem like special categories of data?", *Data Protection Leader*, March 2018, pp. 6-8.

HUME, K., "When is it important for an algorithm to explain itself?", *Harvard Business Review*, 6th July 2018. Available on 16th May 2020 at: https://hbr.org/

HUNT, C. J., FASOLI, F., CARNAGHI, A. & CADINU, M., "Masculine self-presentation and distancing from femininity in gay men: an experimental examination of the role of masculinity threat", *Psychology of Men & Masculinity*, 2016, vol. 17, No. 1, pp. 108-112.

ILLINGWORTH, A. J., "Big data in I-O psychology: privacy consideration and discriminatory algorithms", *Industrial and Organizational Psychology,* vol. 8, No. 4, pp. 567-575.

JEFFREYS, S., "Kate Millett's sexual politics: 40 years on", *Women's Studies International Forum*, vol. 34, 2011, pp. 76-84.

JONES HAVARD, C., "'On the take': the black box of credit scoring and mortgage discrimination", *Public interest law journal*, vol. 20, 2011, pp. 241-287.

JONES, M. L., "The right to a human in the loop: Political constructions of computer automation and personhood", *Social Studies of Science,* vol. 47, No. 2, 2017, pp. 216-239.

KAMARA, I., "Co-regulation in EU personal data protection: the case of technical standards and the privacy by design standardisation 'mandate'", *European Journal of Law and Technology*, vol. 8, No. 1, 2017, pp. 1-25.

KAMINSKI, M. E., "The right to explanation, explained", *Berkeley Technology Law Journal*, vol. 34, No. 1, 2019a, pp. 189-218.

- "Binary governance: Lessons from the GDPR's approach to algorithmic accountability", *Southern California Law Review*, vol. 92, No. 6, 2019b, pp. 1529-1616.

KEHL, D., GUO, P. & KESSLER, S., "Algorithms in the criminal justice system: assessing the use of risk assessments in sentencing", *Responsive Communities Initiative, Berkman Klein Center for Internet & Society, Harvard Law School*, 2017.

KEREN, H., "We insist! Freedom now: Does contract doctrine have anything constitutional to say?", *Michigan Journal of Race Law*, vol. 11, 2005, pp. 133-193.

KIRIMI, J. M. & MOTURI, C. A., "Application of data mining classification in employee performance prediction", *International Journal of Computer Applications*, vol. 146, No. 7, 2016.

KIM, P. T., "Data-driven discrimination at work", *William & Mary Law Review*, vol. 58, 2017, pp. 857-936.

KIM, N. S. & TELMAN, D. A., "Internet giants as quasi-governmental actors and the limits of contractual consent", *Missouri Law Review*, vol. 80, No. 3, 2015, pp. 723-770.

KITZMILLER, E. M., "IDS case study: Allegheny County. Allegheny County's data warehouse: leveraging data to enhance human service programs and policies", *Actionable Intelligence for Social Policy*, May 2014. Available on 4th April 2019 at: https://www.aisp.upenn.edu/

KLEINBERG, J., MULLAINATHAN, S., & RAGHAVAN, M., "Inherent trade-offs in the fair determination of risk scores", 2016. Available on 18th February 2019 at: https://arxiv.org/

KLEINBERG, J., LAKKARAJU, H., LESKOVEC, J., LUDWIG, J. & MULLAINATHAN, S., "Human decisions and machine predictions", *The Quarterly Journal of Economics*, vol. 133, No. 1, 2017, pp. 237-293.

KLEINBERG, J., LUDWIG, J., MULLAINATHAN, S., & SUNSTEIN, C. R., "Discrimination in the age of algorithms", *Journal of Legal Analysis*, vol. 10, 2018, pp. 113-174.

KLONICK, K., "The Facebook oversight board: creating an independent institution to adjudicate online free expression", *Yale Law Journal*, vol. 129, No. 8, 2020, pp. 2418-2499.

KOKOLAKIS, S., "Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon", *Computers & Security*, vol. 107, 2017, pp. 122-134.

KOKOTT, J., & SOBOTTA, C., "The distinction between privacy and data protection in the jurisprudence of the CJEU and the ECtHR", *International Data Privacy Law*, vol. 3, No. 4, 2013, pp. 222-228.

KOOPS, B. J., "The problem with European data protection law", *International Data Privacy Law*, vol. 4, No. 4, 2014, pp. 250-261.

KRAYNAYA, D., "Economics and jurisprudence: is John Rawls' difference principle just another form of supply side economics and can it be applied effectively in modern society?", *Manchester Student Law Review*, vol. 1, 2012, pp. 49-59.

KROLL, J., HUEY, J., BAROCAS, S., FELTEN, E. W., REIDENBERG, J. R., ROBINSON, D. G. & YU, H., "Accountable algorithms", *University of Pennsylvania Law Review*, vol. 165, No. 3, 2017, pp. 633-705.

KÜLLMANN, M., "Platform work, algorithmic decision-making, and EU gender equality law", *The International Journal of Comparative Labour Law and Industrial Relations*, vol. 34, 2018, pp. 1-21.

LANEY, D., "3D data management: controlling data volume, velocity and variety", 6th February 2001. Available on 20th September 2018 at: https://blogs.gartner.com/

LEE, E., "Recognizing rights in real time: the role of Google in the EU right to be forgotten", *University of California Davis Law Review*, vol. 49, No. 3, 2016, pp. 1017-1095.

LEE, J. G., JUN, S., CHO, Y. W., LEE, H., KIM, G. B., SEO, J. B. & KIM, N., "Deep learning in medical imaging: general overview", *Korean Journal of Radiology*, vol. 18, No. 4, 2017, pp. 570-584.

LEHR, D. & OHM, P., "Playing with the data: what legal scholars should learn about machine learning", *UC Davis Law Review*, vol. 51. No. 2, 2017, pp. 653-717.

LEDIEN, J., SOUV, K., LEANG, R., HUY, R., COUSIEN, A., PEAS, M., FROEHLICH, Y., DUBOZ, R., ONG, S., DUONG, V., BUCHY, P., DUSSART, P. & TARANTOLA, A., "An algorithm applied to national surveillance data for the early detection of major dengue outbreaks in Cambodia", *PLOS One*, vol. 14, No. 2, 2019, pp. 1-11.

LERMAN, J., "Big data and its exclusions", *Stanford Law Review Online,* No. 66, 2013, pp. 55-63.

LOFTUS, J., RUSSELL, C., KUSNER, M. J., & SILVA, R., "Causal reasoning for algorithmic fairness", 2018. Available on 30th June 2020 at: https://arxiv.org/

LOJA, E., COSTA, M. E., HUGHES, B. & MENEZES, I., "Disability, embodiment and ableism: stories of resistance", *Disability & Society*, vol. 28, No. 2, 2013, pp. 190-203.

LOUIZOS, C., SWERSKY, K., LI, Y., WELLING, M. & ZEMEL, R., "The variational fair autoencoder", 2015. Available on 8th February 2020 at: https://arxiv.org/

LOWRY, S. & MACPHERSON, G., "A blot on the profession", *British Medical Journal*, 5th March 1988, pp. 657-658.

LUDVIG, A., "Differences between women? Intersecting voices in a female narrative", European Journal of Women's Studies, vol. 13, No. 3, 2006, pp. 245-258.

MACKINNON, R. K., "Pathologising trans people: Exploring the roles of patients and medical personnel", *Theory in Action*, vol. 11, No. 4, 2018, pp. 74-96.

MACNISH, K., "Unblinking eyes: the ethics of automating surveillance", *Ethics and Information Technology*, vol. 14, No. 2, 2012, pp. 151-167.

MAHALINGAM, R., BALAN, S. & HARITATOS, J., "Engendering immigrant psychology: an intersectionality perspective", *Sex Roles*, vol. 59, 2008, pp. 326-336.

MALGIERI, G., "Automated decision-making in the EU Member States: The right to explanation and other 'suitable safeguards'", *Computer Law & Security Review*, vol. 35, No. 5, 2019.

MALGIERI, G. & COMANDÉ, G., "Sensitive-by-distance: quasi-health data in the algorithmic era", *Information & Communications Technology Law*, vol. 26, No. 3, 2017a, pp. 229-249.

- "Why a right to legibility of automated decision-making exists in the General Data Protection Regulation", *International Data Privacy Law*, vol. 7, No. 4, 2017b, pp. 243-265.

MALISZEWSKA-NIENARTOWICZ, J., "Direct and indirect discrimination in European Union law –how to draw a dividing line?", *International Journal of Social Sciences*, vol. 3, No. 1, 2014, pp. 41-55.

MARTÍN CORRALES, E., "Maurofobia/islamofobia y maurofilia/islamofilia en la España del siglo XXI", *Revista CIDOB d'Afers Internacionals,* No. 66-67, 2004, pp. 39-51.

MAS BADÍA, M. D., "Los ficheros de solvencia patrimonial en la proyectada nueva Ley Orgánica de Protección de Datos de carácter personal. ¿Un avance o una oportunidad perdida?", *Actualidad Civil*, No. 11, 2017, pp. 90-112.

MATSUDA, M. J., "Liberal jurisprudence and abstracted visions of human nature: a feminist critique of Rawls' theory of justice", *New Mexico Law Review*, vol. 16, No. 3, 1986, pp. 613-630.

MATTEAZZI, E., PAILHÉ, A. & SOLAZ, A., "Does part-time employment widen the gender wage gap? Evidence from twelve European countries", *Society for the Study of Economic Inequality*, Working Paper 2013-2913, 2013, pp. 1-48.

MAZUR, J., "Automated decision-making and the precautionary principle in EU law", *Baltic Journal of European Studies Tallinn University of Technology*, vol. 9, No. 4, 2019, pp. 3-18.

MELERO GUERVÓS, J. J. & MERELO MOLINA, C., "Evolución de la matrícula femenina en el grado de informática en universidades públicas españolas", 2017. Available on 26th April 2019 at: https://www.researchgate.net/

MESTRE DELGADO, J. F., "Una reflexión sobre la regulación constitucional del Derecho administrativo", *Corts: Anuario de Derecho Parlamentario,* No. Extra 31, 2018, pp. 367-386.

- MESTRE DELGADO, J. F., "El tratamiento jurídico de la discapacidad en la ley de contratos del sector público", *Anales de Derecho y Discapacidad*, No. 3, junio 2018, pp. 87- 99.

MILLER, A. P., "Want less-biased decisions? Use algorithms", *Harvard Business Review*, 26th July 2018. Available on 13th February 2019 at: https://hbr.org/

MIRÓ-LLINARES, F., "Predictive policing: utopia or dystopia? On attitudes towards the use of big data algorithms for law enforcement", *Revista de Internet, Derecho y Política*, No. 30, 2020, pp. 1-18.

MONASTERIO ASTOBIZA, A., "Ética algorítmica: Implicaciones éticas de una sociedad cada vez más gobernada por algoritmos", *Dilemata*, No. 24, 2017, pp. 185-217.

MORAN, R., "Whatever happened to racism?", *St John's Law Review*, vol. 75, No. 4, 2005, pp. 899-927.

MORGENROTH, T. & RYAN, M. K., "Quotas and affirmative action: Understanding group-based outcomes and attitudes", *Social and Personality Psychology Compass*, vol. 12, No. 3, 2018, pp. 1-14.

MORRIS, A. J., "On the normative foundations of indirect discrimination law: Understanding the competing models of discrimination law as Aristotelian forms of justice", *Oxford Journal of Legal Studies*, vol. 15, No. 2, 1995, pp. 199-228.

MITTELSTADT, B. D., ALLO, P., TADDEO, M., WACHTER, S. & FLORIDI, L., "The ethics of algorithms: mapping the debate", *Big Data & Society*, July-December 2016, pp. 1-21.

MUXÍ MARTÍNEZ, Z., CASANOVAS, R., CIOCOLETTO, A., FONSECA, M. & GUTIÉRREZ VALDIVIA, B., "¿Qué aporta la perspectiva de género al urbanismo?" *Feminismo/s*, No. 17, 2011, pp. 105-129.

NAGOSHI, J. L., ADAMS, K. A., TERRELL, H. K, HILL, E. D., BRZUZY, S. & NAGOSHI, C. T, "Gender differences in correlates of homophobia and transphobia", *Sex Roles*, vol. 59, 2008, pp. 521-531.

NEDELSKY, J., "Reconceiving autonomy: sources, thoughts and possibilities", *Yale Journal of Law & Feminism*, vol. 1, 1989, pp. 7-36.

NEFF, G. & NAGY, P., "Talking to bots: symbiotic agency and the case of Tay", *International Journal of Communication*, vol. 10, 2016, pp. 4915–4931.

NELSON, A., "Unequal treatment: confronting racial and ethnic disparities in health care", *Journal of the National Medical Association*, vol. 94, No. 8, 2002, pp. 666-668.

NEUMARK, D., BURN, I. & BUTTON, P., "Is it harder for older workers to find jobs? New and improved evidence from a field experiment", *Journal of Political Economy*, vol. 127, No. 2, 2019, pp. 922-970.

NICHOLSON PRICE II, W., KAMINSKI, M. E., MINSSEN, T., SPECTOR-BAGDADY, K., "Shadow health records meet new data privacy laws", *Science*, vol. 363, No. 6426, 2019, pp. 448-450.

NOGUEIRA, A., "La termita Bolkestein", *El Cronista del Estado Social y Democrático de Derecho*, 2011, pp. 58-70.

NORMAN, P., "Statistical discrimination and efficiency", *The Review of Economic Studies*, vol. 70, No. 3, 2003, pp. 615-627.

NUNZIATO, D. C., "Gender equality: states as laboratories", *Virginia Law Review*, vol. 80, No. 4, 1994, pp. 945-977.

NUSSBAUM, M. C., "Objectification", *Philosophy and Public Affairs*, vol. 24, No. 4, 1995, pp. 249-291.

OHM, P., "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization", *UCLA Law Review*, vol. 57, 2010, pp. 1701-1777.

OKIN, S. M., "Gender, justice and gender: an unfinished debate", *Fordham Law Review*, vol. 72, No. 5, 2004, pp. 1537-1567.

OQUENDO, M.A., BACA-GARCIA, E., ARTÉS-RODRÍGUEZ, A., PÉREZ-CRUZ, F., GALFALVY, H.C., BLASCO-FONTECILLA, H., MADIGAN, D., & DUAN, N., "Machine learning and data mining: strategies for hypothesis generation", *Molecular Psychiatry*, vol. 17, No. 10, 2012, p. 956-959.

ORTEGA ARJONILLA, E. & PLATERO MÉNDEZ, R. E., "Movimientos feministas y trans* en la encrucijada: aprendizajes mutuos y conflictos productivos", *Quaderns de Psicologia*, vol. 17, No. 3, 2015, pp. 17-30.

OSWALD, M., "Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power", *Philosophical Transactions of the Royal Society of London*, *Series A: Mathematical and Physical Sciences*, vol. 376, No. 2128, 2018, pp. 1-20.

OSWALD, M. & GRACE, J., "Intelligence, policing and the use of algorithmic analysis: a freedom of information-based study", *Journal of Information Rights, Policy and Practice*, vol. 1, No. 1, 2016.

OSWALD, M., GRACE, J., URWIN, S. & BARNES, G. C., "Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality", *Information & Communications Technology Law*, vol. 27, No. 2, 2018, pp. 223-250.

PALMA ORTIGOSA, A., "Decisiones automatizadas en el RGPD. El uso de algoritmos en el contexto de la protección de datos", *Revista General de Derecho Administrativo*, No. 50, 2019.

PAUL, R. & HUBER, M., "Risk-based regulation in continental Europe? Explaining the corporatist turn to risk in German work safety policies", *European Policy Analysis*, vol. 1, No. 2, 2015, pp. 5-33.

PARASURAMAN, R. & MILLER, C. A., "Trust and etiquette in high-criticality automated systems", *Communications of the ACM*, vol. 47, No. 4, 2004, pp. 51-55.

PONCE SOLÉ, J., "Inteligencia artificial, Derecho administrativo y reserva de humanidad: algoritmos y procedimiento administrativo debido tecnológico", *Revista General de Derecho Administrativo*, No. 50, 2019, pp. 1-52.

POYANCO BUGUEÑO, R. A., "Los derechos sociales y la libertad: un análisis problemático", *Derecho Público Iberoamericano*, No. 9, 2016, pp. 41-79.

PRZEWORSKI, A., "Conquered or Granted? A History of Suffrage Extensions", *British Journal of Political Science*, vol. 39, No. 2, 2009, pp. 291-321.

PUNYANUT-CARTER, N. M., "The perceived realism of African American portrayals on television", *Howard Journal of Communications*, vol. 19, No. 3, 2008, pp. 241-257.

QUIJANO-SÁNCHEZ, L., LIBERATORE, F., CAMACHO-COLLADOS, J. & CAMACHO-C OLLADOS, M., "Applying automatic text-based detection of deceptive language to police reports: extracting behavioral patterns from a multi-step classification model to understand how we lie to the police", *Knowledge-Based Systems*, vol. 149, 2018, pp. 155-168.

RANA, J., "The story of Islamophobia", *Souls*, vol. 9, No. 2, 2007, pp. 148-161.

RANCHORDÁS, S., "Nudging citizens through technology in smart cities", *International Review of Law, Computers & Technology,* vol. 33, 2019, pp. 1-23.

RANCHORDÁS, S. & SCHUURMANS, Y., "Outsourcing the welfare state: the role of private actors in welfare fraud investigations", *European Journal of Comparative Law and Governance*, vol. 7, No. 2, 2020, pp. 5-42.

RANKIN, K. N., "A critical geography of poverty finance", *Third World Quarterly,* 2013, pp. 547-568.

REIFF, M. M., "The difference principle, rising inequality, and supply-side economics: how rawls got hijacked by the right", *Revue de philosophie économique*, vol 13, No. 2, 2012, pp. 119-173.

RENAN BARZILAY, A. & BEN-DAVID, A., "Platform inequality: gender in the gig economy", *Seton Hall Law Review,* vol. 47, No. 393, 2017, pp. 393-431.

RENDA, A., "Ethics, algorithms and self-driving cars – a CSI of the 'trolley problem'", *CEPS Policy Insights,* No. 2018/02, 2018, pp. 1-15.

REVUELTA PÉREZ, I., "Mejores técnicas disponibles": un singular sistema de regulación ambiental", *Revista Catalana de Dret Ambiental*, vol. 10, No. 1, 2019, pp. 1 -34.

RICHARDS, A. J. & POLAVIEJA, J., "Trade unions, unemployment and working class fragmentation in Spain", *Instituto Juan March de Estudios e Investigaciones*, 1997, pp. 1-57.

RITTER. N., "Predicting recidivism risk: new tool in Philadelphia shows great promise", *National Institute of Justice Journal*, No. 271, 2013, pp. 4-13.

ROBERTS, J. L., "Protecting privacy to prevent discrimination", *William & Mary Law Review*, vol. 56, No. 6, 2015, pp. 2097-2174.

RODRIGUES, R., BARNARD-WILLS, D., DE HERT, P. & PAPAKONSTANTINOU, V., "The future of privacy certification in Europe: an exploration of options under article 42 of the GDPR", *International Review of Law, Computers & Technology*, vol. 30, No. 3, 2016, pp. 248-270.

RODRÍGUEZ RUIZ, B., "¿Identidad o autonomía? La autonomía relacional como pilar de la ciudadanía democrática", *Identidad, Derecho y Política*, No. 17, 2013, pp. 75-104.

RODRÍGUEZ RUÍZ, B. & RUBIO MARÍN, R., "Constitutional justification of parity democracy", *Alabama Law Review*, vol. 60, No. 5, 2009, pp. 1171-1195.

RUBIO, A., "Las políticas de igualdad: de la igualdad formal al mainstreaming", 2003. Available on 15th March 2020 at: http://pmayobre.webs.uvigo.es/

ROSENFELD, M., "Affirmative action, justice, and equalities: a philosophical and constitutional appraisal", *Ohio State Law Journal*, vol. 46, No. 4, 1985, pp. 845-924.

ROTHSTEIN, H., DEMERITT, D., PAUL, R., BEAUSSIER, A., WESSELING. M., HOWARD, M., DE HAAN, M., BORRAZ, O., HUBER, M. & BOUDER, F., "Varieties of risk regulation in Europe: coordination, complementarity and occupational safety in capitalist welfare states", *Socio-Economic Review*, vol. 17, No. 4, 2017, pp. 993-1020.

RUBIO, A., "La eficacia de la legislación española en materia de igualdad de género", *Gênero & Direito*, vol. 4, No. 1, 2015

RUDDMAN, D., L., "Shaping the active, autonomous and responsible modern retiree: an analysis of discursive technologies and their links with neo-liberal political rationality", *Ageing & Society*, No. 26, 2006, pp. 181-201.

RUIZ-CANTERO, M. T., VIVES-CASES, C., ARTAZCOZ, L., DELGADO, A., GARCÍA CALVENTE, M. D. M., MIQUEO, C., MONTERO, I., ORTIZ, R., RONDA, E., RUÍZ, I. & VALLS, C. "A framework to analyse gender bias in epidemiological research", *Journal of Epidemiology and Community Health*, vol. 61, No. 2, 2007, pp. 46-53.

SAHU, H., SHRMA S., & GONDHALAKAR, S., "A brief overview on data mining survey", *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, vol. 1, No. 3, 2013, pp. 114-121.

SALOMÉ, L. M., "La discriminación y algunos de sus calificativos: directa, indirecta, por indiferenciación, interseccional (o múltiple) y estructural", *Revista de Pensamiento Constitucional*, vol. 22, No. 22, 2017, pp. 255-290.

SÁNCHEZ-MARTÍNEZ, F. I., ABELLÁN-PERPIÑÁN, J. M. & OLIVA-MORENO, J., "Privatization in healthcare management: an adverse effect of the economic crisis and a symptom of bad governance. SESPAS report 2014", *Gaceta Sanitaria*, vol. 28, No. 1, 2014, pp. 75-80.

SAURWEIN, F., JUST, N. & LATZER, M., "Governance of algorithms: options and limitations", *Digital Policy, Regulation and Governance*, vol. 17, No. 6, 2015, pp. 35-49.

SCHARTUM, W., "Law and algorithms in the public domain", *Etikk i praksis, Nordic Journal of Applied Ethics*, No. 1, 2016, pp. 15-26.

SCHERER, M., "Regulating artificial intelligence systems: risks, challenges, competencies, and strategies", *Harvard Law & Technology Journal*, vol. 29, No. 2, 2016, pp. 353-400.

SCHERMER, B. W., "The limits of privacy in automated profiling and data mining", *Computer Law & Security Review*, vol. 27, No. 1, 2011, pp. 45-52.

SCHREER, G. E., SMITH, S. & THOMAS, K., "'Shopping while black': examining racial discrimination in a retail setting", *Journal of Applied Social Psychology*, vol. 39, No. 6, 2009, pp. 1432-1444.

SCHLEUTKER, E., "Discrimination against religious minorities", *Journal of Church and State*, 2018, pp. 1-26.

SERRANO PARTIDA, R., "El desafío catalán y el fin de la transición democrática", *Razón y Palabra*, vol. 22, No. 100, 2018, pp. 83-92.

SILVERS, A. & STEIN, M. A., "Disability and the social contract", *The University of Chicago Law Review,* vol. 74, 2007, pp. 1615-1635.

SKITKA, L. J., MOSIER, K. L, BURDICK, M. & ROSENBLATT, B., "Automation bias and errors: are crews better than individuals?", *The International Journal of Aviation Psychology*, vol. 10, No. 1, 2000, pp. 85-97.

STAUFFER, J. M. & BUCKLEY, M. R., "The existence and nature of racial bias in supervisory ratings", *Journal of Applied Psychology*, vol. 90, No. 3, 2005, pp. 586-591.

STEPAN, N. L., "Race, Gender, Science and Citizenship", *Gender & History*, vol. 10, No. 1, 1998, pp. 26-52.

STRAHILEVITZ, L., "Privacy versus antidiscrimination", *University of Chicago Law Review*, vol. 75, No.1, 2008, pp. 363-381.

SULLIVAN, M. K., "Homophobia, history, and homosexuality", *Journal of Human Behavior in the Social Environment,* vo. 8, No. 2-3, 2004, pp. 1-13.

SUNSTEIN, C., "Sludge and ordeals", *Duke Law Review*, vol. 68, No. 8, 2018, pp. 1843-1883.

SURDEN, H., "Machine learning and law", *Washington Law Review,* vol. 89, 2014, pp. 87-115.

SWEENEY, L., "Discrimination in online ad delivery", *Communications of the ACM*, vol. 56, No. 5, 2013, pp. 44-54.

THOMAS, V. & AZMITIA, M., "Does class matter? The centrality and meaning of social class identity in emerging adulthood", *Identity,* vol. 14, No. 3, pp. 195-213.

TODOLÍ SIGNES, A., "La gobernanza colectiva de la protección de datos en las relaciones laborales: big data, creación de perfiles, decisiones empresariales automatizadas y los derechos colectivos", *Revista de Derecho Social*, 2018, No. 84, pp. 69-88.

TONER, H., "Impact assessments and fundamental rights protection in EU law", *European Law Review*, No. 3, 2006, pp. 316-341.

TUTT, A., "An FDA for algorithms", *Administrative Law Review*, vol. 69, No. 1, 2017 pp. 83-123.

TVERSKY, A. & KAHNEMAN, D., "Judgment under uncertainty: heuristics and biases", *Science*, vol. 147, No. 4157, 1974, pp. 1124-1131.

VALLETTI, T., & WU, J., "Consumer profiling with data requirements", *Production and Operations Management*, Vol. 29, No. 2, 2020, pp. 309-329.

VEDDER, A. & NAUTS, L., "Accountability for the use of algorithms in a big data environment", *International Review of Law, Computers & Technology*, vol. 31, No. 2, pp. 206-224.

VERLOO, M., "Multiple inequalities, intersectionality and the European Union", *European Journal of Women's Studies*, vol. 13, No. 3, 2006, pp. 211-228.

WANG, J., MA, Y., ZHANG, L., GAO, R. X. & WU, D., "Deep learning for smart manufacturing: Methods and applications", *Journal of Manufacturing Systems*, vol. 48, 2018, pp. 144-156.

WARD, J., & BARKER, A., "Undefined by data: a survey of big data definitions", 2013. Available on 20th September 2018: https://arxiv.org/

WACHTER, S., MITTELSTADT, B. D. & FLORIDI, L., "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation", *International Data Privacy Law,* vol. 7, No. 2, 2017, pp. 76-99.

WACHTER, S., MITTELSTADT, B. D. & RUSSELL, C., "Counterfactual explanations without opening the black box: automated decisions and the GDPR", *Harvard Journal of Law & Technology*, vol. 31, No. 2, 2018, pp. 841-887.

WACHTER, S. & MITTELSTADT, B. D., "A right to reasonable inferences: re-thinking data protection law in the age of big data and AI", *Columbia Business Law Review*, vol. 2019, No. 2, 2019, pp. 494-620.

WACHTER, S., "Affinity profiling and discrimination by association in online behavioural advertising", *Berkeley Technology Law Journal*, vol. 35, No. 2, 2020 (forthcoming), pp. 1-74. Available on 15th March 2020 at: https://papers.ssrn.com/

WESTWOOD, S. J., IYENGAR, S., WALGRAVE, S., LEONISIO, R., MILLER, L. & STRIJBIS, O., "The tie that divides: Cross-national evidence of the primacy of partyism", *European Journal of Political Research*, vol. 57, 2018, pp. 333-354.

WHEELER, S. C., JARVIS, W., B., G., PETTY, R., E., "Think unto others: the self-destructive impact of negative racial stereotypes", *Journal of Experimental Social Psychology*, vol. 37, No. 2, 2001, pp. 173-180.

WHITLEY JR., B. E., "Gender-role variables and attitudes toward homosexuality", *Sex roles*, vol. 45, No. 11/12, 2001, pp. 691-721.

WIEDEMANN, K., "Automated processing of personal data for the evaluation of personality traits: legal and ethical issues", *Max Planck Institute for Innovation and Competition Research Paper No. 18-04*, 2018. Available on 29th July 29 2019 at: https://ssrn.com/

WILK, A. “Teaching AI, ethics, law and policy”, 24th May 2019, pp. 1-7. Available on 15th June 2019 at: https://arxiv.org/

WILLIAMS, B. A., BROOKS, C. F. & SHMARGAD, Y., “How algorithms discriminate based on data they lack: challenges, solutions, and policy implications”, *Journal of Information Policy*, No. 8, 2018, pp. 78-115.

WINTERS, N., EYNON, R., GENIETS, A., ROBSON, J. & KAHN, K., “Can we avoid digital structural violence in future learning systems?”, *Learning, Media and Technology*, vol. 45, No. 1, 2020, pp. 17-30.

WRIGHT, P. J. & TOKUNAGA, R. S., “Men’s objectifying media consumption, objectification of women, and attitudes supportive of violence against women”, *Archives of Sexual Behavior*, vol. 45, No. 4, 2016, pp. 955-964.

YANISKY-RAVID, S. & HALLISEY, S., “Equality and privacy by design: a new model of artificial intelligence data transparency via auditing, certification, and safe harbor regimes”, *Fordham Urban Law Journal*, vol. 46, No. 2, 2019, pp. 428-486.

YEUNG, K., “‘Hypernudge’: Big Data as a mode of regulation by design”, *Information, Communication & Society*, vol. 20, No. 1, 2017, pp. 118-136.

YEUNG, K., “Algorithmic regulation: a critical interrogation”, *Regulation & Governance*, vol. 12, No. 4, 2018, pp. 505-523.

YOUNG, E. “Educational privacy in the online classroom: FERPA, MOOCs, and the Big Data Conundrum", *Harvard Journal of Law & Technology*, vol. 28, No. 2, 2015, pp. 549-592.

ZARSKY, T., “Transparent Predictions”, *Illinois Law Review*, vol. 2013, No. 4, 2013, pp. 1503-1570.

- “Understanding discrimination in the scored society”, *Washington Law Review*, vol. 89, No. 4, 2014, pp. 1375-1412.

- “The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making”, *Science, Technology, and Human Values*, vol. 41, No. 1, 2016, pp. 118-132.

- “Incompatible: the GDPR in the age of big data”, *Seton Hall Review*, vol. 47, 2017, pp. 995-1020.

ZERILLI, J., KNOTT, A., MACLAURIN, J. & GAVAGHAN, C., “Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?”, *Philosophy and Technology*, 2018, pp. 1-23.

ŽLIOBAITĖ, I., “A survey on measuring indirect discrimination in machine learning”, September 2015. Available on 25th September 2018 at: https://arxiv.org/

ŽLIOBAITĖ, I. & CUSTERS, B., "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models", *Artificial Intelligence & Law*, vol. 24, No. 2, 2016, pp. 183-201.

ZUIDERVEEN BORGESIUS, F., & POORT, J., "Online Price Discrimination and EU Data Privacy Law," *Journal of Consumer Policy*, vol. 40, 2017, pp. 347-366.

ZUIDERVEEN BORGESIUS, F., "Price discrimination, algorithmic decision-making, and european non-discrimination law", *European Business Law Review* (Forthcoming), 2020, pp. 1-29. Available on 10th June 2020 at: https://ssrn.com/

ŻUK, P., "One leader, one party, one truth: public television under the rule of the populist right in Poland in the pre-election period in 2019, *Javnost – The Public, Journal of the European Institute for Communication and Culture*, vol. 27, No. 3, 2020, pp. 287-307.

2. Speeches and papers presented at conferences and proceedings

BAROCAS, S., "Data mining and the discourse on discrimination", *Proceedings of the Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining (KDD)*, 2014. Available, on 20th February 2019 at: https://pdfs.semanticscholar.org/

BINNS, R., VAN KLEEK, M., VEALE, M., LYNGS, U., ZHAO, J. & SHADBOLT, N., "It's Reducing a Human Being to a Percentage; Perceptions of Justice in Algorithmic Decisions", paper presented at the *ACM Conference on Human Factors in Computing Systems (CHI'18*), 2018. Available on 10th May 2019 at: https://arxiv.org/abs/

DATTA, A., TSCHANTZ, M. C. & DATTA, A., "Automated experiments on ad privacy settings: a tale of opacity, choice and discrimination", *Proceedings on Privacy Enhancing Technologies 2015*, 2015, pp. 92-112.

DE-ARTEAGA, M. FLOGLIATO, R. & CHOULDECHOVA, A., "A case for humans-in-the-loop: decisions in the presence of erroneous algorithmic scores", in *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, Association for computing machinery, 2020, pp. 1-12.

DE MAURO, A., GRECO, M. & GRIMALDI, M., "What is big data? A consensual definition and a review of key research topics", paper presented at the *4th International Conference on Integrated Information*, Madrid, 5-8th September 2015, pp. 97-104.

DWORK, C., HARDT, M., PITASSI, T., REINGOLD, O. & ZEMEL, R., "Fairness through awareness", *ITCS '12 Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012, pp. 214-226.

FALIAGKA, E., RAMANTAS, K. & TSAKALIDIS, A., "Application of machine learning algorithms to an online recruitment system", paper presented at the *7th International Conference on Internet and Web Applications and Services*. Available on 26th March 2019 at: http://citeseerx.ist.psu.edu/

FERREIRA, G., "Software certification in practice: how are standards being applied?", paper presented at the *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*, Buenos Aires, 2017, pp. 100-102.

FISHER, A., MARGOLIS, J. & MILLER, F., "Undergraduate women in computer science: experience, motivation and culture", *SIGCSE '97 Proceedings of the twenty-eighth SIGCSE technical symposium on Computer science education*, 1997, pp. 106-110.

FREIXAS, L., "La mujer callada de todos es alabada", speech delivered at the *IV Conference Feminario*, Valencia, 2019. Available on 10th April 2019 at: https://www.youtube.com/

FELDMAN, M., FRIEDLER, S. A., MOELLER, J., SCHEIDEGGER, C. & VENKATASUBRAMANIAN, S., "Certifying and removing disparate impact", *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 259-268.

GALLEGOS ARGÜELLO, M. C., "La identidad de género: masculino versus femenino", in SUÁREZ VILLEGAS, J. C., LIBERIA VAYÁ, I. & ZURBANO-BERENGUER, B. (coords.), *I Congreso Internacional de Comunicación y Género. Libro de Actas: 5, 6 y 7 de marzo de 2012. Facultad de Comunicación. Universidad de Sevilla*, Sevilla, Universidad de Sevilla, Facultad de Comunicación, 2012, pp. 705-718.

GARIBÓ I ORTS, O., "Multilingual detection of hate speech against immigrants and women in twitter at semeval-2019 task 5: frequency analysis interpolation for hate in speech detection", *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, pp. 460-463.

GOODMAN, B. W., "A step towards accountable algorithms?: Algorithmic discrimination and the European Union general data protection", paper presented at the *29th Conference on Neural Information Processing Systems*, Barcelona, 2016. Available on 13th February 2019 at: http://www.mlandthelaw.org/

HARDT, M., PRICE, E. & SREBRO, N., "Equality of Opportunity in Supervised Learning", paper presented at the *30th Conference on Neural Information Processing Systems*, Barcelona, 2016. Available on 2nd July 2020 at: http://papers.nips.cc/

JOOSTEN, J., "Control funcional y control cualitativo de los algoritmos en la administración pública", paper presented at the II Seminario internacional DAIA, 10th and 11th October 2019. Slides available on 6th July 2020 at: https://www.phil.uu.nl/

LIU, H. & MOTODA, H., "Feature selection, extraction and construction", paper presented at the *Towards the Foundation of Data Mining Workshop*, *Sixth Pacific – Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2002)*, Taipei, Taiwan, 2002, pp. 67-72.

LUSTIG, N., ARIAS, O., RIGOLINI, J., "Reducción de la pobreza y crecimiento económico: la doble casualidad", paper presented at the *Seminario 'La Teoría del Desarrollo en los Albores del Siglo XXI'*, 2002.

SHNEIDERMAN, B., "Algorithmic Accountability: Designing for safety through human-centered independent oversight", *Turing Lecture (The Alan Turing Institute)*, 31st May 2017. Available on 23rd May 2020 at: https://www.youtube.com/

SPEICHER, T., ALI, M., VENKATADRI, G., RIBEIRO, F. N., ARVANITAKIS, G., BENEVENUTO, F., GUMMADI, K. P, LOISEAU, P. & MISLOVE, A., "Potential for discrimination in online targeted advertising", *Proceedings of Machine Learning Research*, 81, 2018, pp. 1-15.

WICK, M., PANDA, S. & TRISTAN, J. B., "Unlocking fairness: a trade-off revisited", paper presented at the conference on *Advances in Neural Information Processing Systems (NIPS 2019)*, 2019. Available on 16th June 2020 at: https://papers.nips.cc/

3. Book chapters

ALEXY, R., "Sobre los derechos constitucionales a protección", in GARCÍA MANRIQUE, R., (ed.), *Derechos Sociales y Ponderación*, Madrid, Fundación coloquio jurídico europeo, 2nd ed., 2009, pp. 45-84.

AÑÓN ROIG, M. J., "Grupos sociales vulnerables y derechos humanos. Una perspectiva desde el derecho antidiscriminatorio", in ANSUÁTEGUI ROIG, J., RODRÍGUEZ URIBES, J. M., PECES-BARBA MARTÍNEZ, G. & FERNÁNDEZ GARCÍA, E., (coords.), *Historia de los Derechos Fundamentales*, Madrid, Dykinson, 2013a, pp. 613-671.

BALL, K., "Blacklists and black holes: credit scoring in Europe", in WEBSTER, W., & BALL, K., *Surveillance and Democracy in Europe: Courting Controversy*, 2019, Oxon, Routledge.

BARRANCO, M. C., "La concepción republicana de los derechos en un mundo multicultural", in DEL REAL ALCALÁ, J. A., ANSUÁTEGUI ROIG, F. J., LÓPEZ GARCÍA, J. A. & RUIZ RUIZ, R., (coords.), *Derechos Fundamentales, Valores y Multiculturalismo*, Madrid, Dykinson, 2005, pp. 15-35.

BARRÈRE UNZUETA, M. A., "Iusfeminismo y derecho antidiscriminatorio: hacia la igualdad por la discriminación", in MESTRE, R., (coord.), *Mujeres, Derechos y Ciudadanías*, Valencia, Tirant Lo Blanch, 2008, pp. 45-72.

BING, J., "Code, access and control" in MURRAY, A., & KLANG, M., *Human Rights in the Digital Age*, London, Glasshouse Press, 2005, pp. 203-2018.

ABATE, M., "How to prevent crimes using earthquakes", in EMMER, M., & ABATE, M., (eds.), *Imagine Math 6: Between Culture and Mathematics*", Switzerland, Springer, pp. 103-114.

CALDERS, T. & CUSTERS, B., "What is data mining and how does it work?", in CUSTERS, B., CALDERS, T., SCHERMER, B. & ZARSKY, T., (eds.), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in large Databases*, Berlin, Springer, 2013, pp. 27-42.

CALDERS, T. & ŽLIOBAITĖ, I., "Why unbiased computational processes can lead to discriminative decision procedures" CUSTERS, B., CALDERS, T., SCHERMER, B., & ZARSKY, T., (eds.), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in large Databases*, Berlin, Springer, 2013, pp. 43-57.

CATE, F. H., "The failure of fair information practice principles", in WINN, J. K., *Consumer Protection in the Age of the Information Economy*, Abingdon, Routledge, 2006, pp. 341-378.

CRIADO, N. & SUCH, J. M., "Digital discrimination", in YEUNG, K. & LODGE, M., (eds.), *Algorithmic Regulation*, Oxford, Oxford University Press, 2019, pp. 82-97.

CUSTERS, B. "Data dilemmas in the information society: introduction and overview", in CUSTERS, B., CALDERS, T., SCHERMER, B., & ZARSKY, T., (eds.), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in large Databases*, Berlin, Springer, 2013, pp. 3-26.

DIMITROVA, D. & DE HERT, P., "The right of access under the police directive: small steps forward", in MEDINA M., MITRAKAS, A., RANNENBERG, K., SCHWEIGHOFER, E. & TSOUROULAS, N., (eds.), *6th Annual Privacy Forum 2018: Privacy Technologies and Policy*, Berlin, Springer, 2018, pp. 111-130.

DONOHUE, J. J., "Antidiscrimination law", in POLINSKY, A. M. & SHAVELL, S., *Handbook on Law and Economics*, vol. 2, Amsterdam, North-Holland (Elsevier), 2007, pp. 1387-1472.

ESMENGER, N., "Making programming masculine", in MISA, T. J., (ed.), *Gender codes: Why Women are Leaving Computing*, Hoboken (New Jersey), John Wiley & Sons, 2010, pp. 115-142.

FERRE, M. M., "Soft Repression: Ridicule, Stigma, and Silencing in Gender-Based Movements", in DAVENPORT, C., JOHNSTON, H. & MUELLER, C., (eds.), *Repression and Mobilization*, Minneapolis, University of Minnesota Press, 2005, pp. 138-155.

FLANAGAN, M., HOWE, D. C., & NISSENBAUM, H., "Embodying values in technology: theory and practice", in VAN DEN HOVEN, J. & WECKETT, J., (eds.), *Information Technology and Moral Philosophy,* Cambridge, Cambridge University Press, 2008, pp. 322-353.

GELLERT, R., DE VRIES, K., DE HERT, P. & GUTWIRTH, S., "A comparative analysis of anti-discrimination and data protection legislations", in CUSTERS, B., CALDERS, T., SCHERMER, B. & ZARSKY, T., (eds.), *Discrimination and Privacy in the Information Society: Data Mining and Profiling in large Databases*, Berlin, Springer, 2013, pp. 61-88.

GRIFFITHS, A., "The practical challenges of implementing algorithmic regulation for public services", in YEUNG, K. & LODGE, M., (eds.), *Algorithmic Regulation*, Oxford, Oxford University Press, 2019, pp. 150-177.

GONZÁLEZ RAMS, P., "Las mujeres con discapacidad y sus múltiples desigualdades; un colectivo todavía invisibilizado en los estados latinoamericanos y en las agencias de cooperación internacional", in REY TRISTÁN, E. & CALVO GONZÁLEZ, P., (coords.), *200 Años de Iberoamérica (1810-2010)*, 2010, pp. 2737-2756.

HEPPLE, B., "Equality at work", in HEPPLE, B. & VENEZIANI, B., *The Transformation of Labour Law in Europe: A Comparative Study of 15 Countries 1945-2004*, Portland, Hart Publishing, 2009, pp. 129-164.

HILDEBRANDT, M., "The dawn of a critical transparency right for the profiling era", in BUS, J., CROMPTON, M., HILDEBRANDT, M. & METAKIDES, G., (eds.), *Digital Enlightenment Yearbook 2012*, Amsterdam, IOS Press, 2012, pp. 41-56.

LIU, J. & WU, C., "Deep learning based recommendation: a survey" in KIM K. & JOUKOV N., (eds), *Information Science and Applications 2017. ICISA 2017*, Lecture Notes in Electrical Engineering, vol. 424, Singapore, Springer, pp. 451-458.

LOOS, E. & IVAN, L., "Visual ageing in the media", in AYALON, L. & TESCH-RÖMER, C. (eds.), *Contemporary perspectives on ageism*, Cham, Springer, 2018, pp. 163-176.

MESTRE DELGADO, J. F., "Artículo 105" in RODRÍGUEZ-PIÑERO Y BRAVO-FERRER, M. & CASAS BAAMONDE, M. E., (dirs.), *Comentarios a la Constitución Española*, *Tomo II*, Madrid, Fundación Wolters Kluwer, Boletín Oficial del Estado, Tribunal Constitucional y Ministerio de Justicia, 2018, pp. 487-501.

SCATAMBURLO, T., CHARLESWORTH, A. & CRISTIANINI, N., "Machine decisions and human consequences", in YEUNG, K. & LODGE, M., (eds.), *Algorithmic Regulation*, Oxford, Oxford University Press, 2019, pp. 49-81.

SCHINZEL, B., "Cultural differences of female enrolment in tertiary education in Computer Science", in BRUNNSTEIN, K. & BERLEUR, J., (eds.), *Human Choice and Computers: Issues of Choice and Quality of Life in the Information Society*, Berlin, Springer, 2002, pp. 283-292.

SCHREURS, W., HILDEBRANDT, M., KINDT, E. & VANFLETEREN, M., "Cogitas, Ergo Sum. The Role of Data Protection Law and Non-discrimination Law in Group Profiling in the Private Sector" in HILDEBRANDT, M. & GUTWIRTH, S. (eds.), *Profiling the European Citizen*, Berlin, Springer, 2008, pp. 241-270.

VEALE, M. & BRASS, I., "Administration by algorithm", in YEUNG, K. & LODGE, M., (Eds.), *Algorithmic Regulation*, Oxford, Oxford University Press, 2019, pp. 121-149.

VELASCO RICO, C. I., "Vigilando al algoritmo. Propuestas organizativas para garantizar la transparencia", in PUENTES COCIÑA, B. & QUINTIÁ PASTRANA, A., (dirs.), *El Derecho ante la Transformación Digital*, Barcelona, Atelier, 2019, pp. 73-89.

WAGNER, B., "Ethics as an escape from regulation: from ethics-washing to ethics-shopping?", in HILDEBRANDT, M., (ed.), *Being profiled. Cogitas Ergo Sum*, Amsterdam, Amsterdam University Press, 2018, pp. 84-89.

YEUNG, K. & LODGE, M., "Algorithmic regulation: an introduction", in YEUNG, K. & LODGE, M., (eds.), *Algorithmic Regulation*, Oxford, Oxford University Press, 2019, pp. 1-18.

YEUNG, K., "Why worry about decision-making by machine?", in YEUNG, K. & LODGE, M., (eds.), *Algorithmic regulation*, Oxford, Oxford University Press, 2019, pp. 21-48.

ZIBUSCHKA, J., KUROWSKI, S., ROßNAGEL, H., SCHMUCK, C. H. & ZIMMERMANN, C., "Anonymization is dead – long live privacy", in ROßNAGEL, H., WAGNER, S. & HÜHNLEIN, D., (eds.), *Open Identity Summit 2019*, Bonn, Gesellschaft für Informatik, 2019, pp. 71-82.

WALKER, A. D. & SMITH, L., "Social class oppression as social exclusion: a relational perspective", in HAMMACK, P. L., *The Oxford Handbook of Social Psychology and Social Justice*, New York, Oxford University Press, 2018, pp. 245-260.

XENIDIS, R. & SENDEN, L., "EU Non-discrimination law in the era of artificial intelligence: mapping the challenges of algorithmic discrimination", in BERNITZ, U., GROUSSOUT, X., PAJU, J. & DE VRIES, S., (eds.), *General Principles of EU Law and the EU Digital Order*, Alphen aan den Rijn, Kluwer Law International, 2020, pp. 151-182.

4. Books

AGUILERA RULL, A., *Contratación y Diferencia: La Prohibición de Discriminación por Sexo y Origen Étnico en el Acceso a Bienes y Servicios*, València, Tirant lo Blanch, 2013.

ALEXY, R., *A Theory of Constitutional Rights*, Translation by Julian Rivers, Oxford, Oxford University Press, 2002.

ATIENZA, M., *El Sentido del Derecho*, Barcelona, Ariel, 7ª ed., 2018.

BARRANCO, M. C., *Diversidad de Situaciones y Universalidad de los Derechos*, Madrid, Dykinson, 2011.

BARRÈRE, M. A., *Discriminación, Derecho Antidiscriminatorio y Acción Positiva a favor de las Mujeres*, Madrid, Civitas, 1997.

BERLIN, I., *Liberty*, Oxford, Oxford University Press, 2002.

BILBAO UBILLOS, J. M., *La Eficacia de los Derechos Fundamentales frente a Particulares: Análisis de la Jurisprudencia del Tribunal Constitucional*, Madrid, Centro de Estudios Políticos y Constitucionales, 1997.

BODDINGTON, P., *Towards a Code of Ethics for Artificial Intelligence*, Oxford, Springer, 2017.

BOSTRON, N., *Superintelligence. Paths, Dangers, Strategies*, Oxford University Press, Oxford, 2014.

BRYANT, G. *The Working Woman Report: Succeeding in Business in the 80's,* New York, Simon & Schuster, 1984.

BUTLER, J., *Gender Trouble*, London, Routledge, 1990.

CALERO, J., (dir.), *Desigualdades Socioeconómicas en el Sistema Educativo Español*, Madrid, Secretaría General Técnica del Ministerio de Educación y Ciencia, 2007.

CONNOLLY, T., & BEGG, C., *Database Systems: A Practical Approach to Design, Implementation, and Management*, Essex, Pearson, 6th ed., 2015.

CORTINA, A., *Aporofobia, el rechazo al pobre*, Barcelona, Paidós, 2017.

COSCULLUELA MONTANER, L., *Manual de Derecho Administrativo*, Cizur Menor, Aranzadi, 27th ed., 2017.

CRAIG, P. & DE BURCA, G., *EU Law: Text Cases and Materials*, New York, Oxford University Press, 5th ed., 2011.

CRAWFORD, C., DEARDEN, L., MICKLEWRIGHT, J. & VIGNOLES, A., *Family Background and University Success: Differences in Higher Education Access and Outcomes in England*, Oxford, Oxford University Press, 2017

CROCKER, L., *Positive Liberty: An Essay in Normative Political Philosophy*, The Hague, Martinus Nijhoff Publishers, 1980.

DARNACULLETA I GARDELLA, M., *Autorregulación y Derecho Público: La Autorregulación Regulada*, Madrid, Marcial Pons, 2005.

DE LORA, P., *Lo Sexual es Político (y Jurídico)*, Madrid, Alianza Editorial, 2019.

DOMÉNECH PASCUAL, G., *Derechos Fundamentales y Riesgos Tecnológicos*, Madrid, Centro de Estudios Políticos y Constitucionales, 2006.

DWORKIN, R., *Sovereign Virtue*, Cambridge (Massachusetts), Harvard University Press, 4th ed., 2002.

DWORKIN, R., *Taking Rights Seriously*, Cambridge (Massachusetts), Harvard University Press, 1977.

ELLIS, E. & WATSON, P., *EU Anti-discrimination Law*, Oxford, Oxford University Press, 2nd ed., 2012.

ESMENGUER, N. L., ASPRAY, W. JR. & MISA, T. J., *Computer Boys Take Over: Computers, Programmers, and the Politics of Technical Expertise*, Cambridge (Massachusetts), MIT Press, 2010.

EUBANKS, V., *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, New York, St Martin's Press, 2017.

FERNÁNDEZ RUÍZ-GÁLVEZ, E., *Igualdad y Derechos Humanos,* Madrid, Tecnos, 2003.

FRANKS, B., *Taming the Big Data Tidal Wave,* Hoboken, New Jersey, John Wiley & Sons, 2012.

FREDMAN, S., *Discrimination Law*, New York, Oxford University Press, 2nd ed., 2011.

GERARDS, J., *Judicial Review in Equal Treatment Cases*, Leiden, Koninklijke Brill NV, 2005.

GIL GONZÁLEZ, E., *Big data, Privacidad y Protección de Datos,* Madrid, Agencia Española de Protección de Datos, 2016.

GONZÁLEZ, F. J., *El Fin del Mito Masculino: La Entrada en el Siglo de la Mujer*, Barcelona, Erasmus Ediciones, 2007.

HAYEK, F., *New Studies in Philosophy, Politics, Economics and the History of Ideas*, London, Routledge, 1978.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Berlin, Springer, 2009.

HOFFMANN-RIEM, W., *Big Data. Desafíos también para el Derecho*, translation by Eduardo Knörr Argote, Cizur Menor, Aranzadi, 2018.

JONES, E. L., *The European Miracle: Environments, Economies and Geopolitics in the History of Europe and Asia*, Cambridge, Cambridge University Press, 1981.

KANT, I., *Grounding for the Metaphysics of Morals*, Indiannapolis, Hackett Publishing Company, 3rd ed, 1993 (first published in 1785).

LERNER, G., *The Creation of Patriarchy*, Oxford, Oxford University Press, 1987.

LESSIG, L., *Code: Version 2.0*, New York, Basic books, 2006.

LIPPI-GREEN, R., *English with an Accent*, London, Routledge, 2nd ed., 2012.

LOCKE, J., *Two Treatises of Government. Second essay: Concerning the True Original Extent and End of Civil Government*. Available on 15th April 2019 at: http://www.yorku.ca/

LYNSKEY, O., *The Foundations of EU Data Protection Law*, Oxford, Oxford University Press, 2015.

MARSDEN, C.T., *Internet Co-regulation: European Law, Regulatory Governance and Legitimacy in Cyberspace*, Cambridge, Cambridge University Press, 2011.

MCRUER, R., *Crip Theory: Cultural Signs of Queernes and Disability*, New York, New York University Press, 2006.

MUÑOZ MACHADO, S., *Tratado de Derecho Administrativo y de Derecho Público General. Tomo XIV. La Actividad Regulatoria de la Administración*, Madrid, Boletín Oficial del Estado, 2015.

MURPHY, K. P., *Machine Learning: A Probabilistic Perspective*, Cambridge (Massachusetts), The MIT Press, 2012.

NAGY, C. I., *Collective Actions in Europe: A Comparative, Economic and Transsystemic Analysis*, Cham, Springer, 2019.

NATHANSON, S., *Economic Justice*, New Jersey, Prentice Hall, 1988. Available on 24th April 2019 at: http://www.woldww.net/

NEWTON, M. & MAY, L. A., *Proportionality in International Law*, Oxford, Oxford University Press, 2014.

NOBLE, S. U., *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, New York University Press, 2018.

NOZICK, R., *Anarchy, State and Utopia*, Oxford, Blackwell, 1974.

NUSSBAUM, M., *Women and Human Development: The Capabilities Approach*, New York, Cambridge University Press, 2000.

OKIN, S. M., *Justice, Gender and the Family*, New York, Basic Books, 1997.

OKUN, A. M., *Equality and Efficiency: The Big Tradeoff*, Washington DC, Brookings Institution Press, 2015 (1st ed. 1975).

O'NEIL, C., *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, London, Penguin Books, 2017.

PASQUALE, F., *The Black Box Society: The Secret Algorithms that Control Money and Information*, Cambridge (Massachusetts), Harvard University Press, 2015.

PATEMAN, C. *The Sexual Contract*, Stanford, Stanford University Press, 1988.

PATEMAN, C. & MILL, C., *Contract & Domination*, Cambridge, Polity Press, 2007.

POLAND, B., *Haters: harassment, abuse and violence online*, Lincoln, Potomac Books, 2016.

RAWLS, J., *A Theory of Justice*, Cambridge (Massachusetts), Harvard University Press, 1971 (original edition), 1999 (revised edition).

ROUSSEAU, J. J., *The Social contract*, 2017 (first published in 1762). Available on 15th April 2019 at: https://www.earlymoderntexts.com/

SEN, A., *The Idea of Justice*, Cambridge (Massachusetts), The Belkap Press of Harvard University Press, 2009.

SUNSTEIN, C. R. Y THALER, R. H., *Nudge: Improving Decisions about Health, Wealth and Happiness*, New Haven, Yale University Press, 2008.

SUTHAHARAN, S., *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, New York, Springer.

TASA FUSTER, V., *Llengua i Estat: Suïssa i Espanya davant la Diversitat Lingüística*, València, Universitat de València, Servei de Publicacions, 2019.

TEGMARK, M., *Life 3.0. Being Human in the Age of Artificial Intelligence*, Penguin Books, London, 2017.

TYLER, T. R., *Why People Obey the Law*, New Haven and London, Yale University Press, 1990.

VANDENHOLE, W., *Non-discrimination and Equality in the View of the UN Human Rights Treaty Bodies*, Antwerpen – Oxford, Intersentia, 2005.

WILLIAMS, E., *Capitalism and Slavery*, Chapel Hill, the University of North Carolina Press, 1944.

WITTEN, I. H., FRANK, E., HALL, M. A. & PAL, C. J., *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Cambridge (Massachusetts), Morgan Kaufman, 2017.

WOLLSTONECRAFT, M., *A Vindication of the Rights of Woman*, 2017 (first published in 1974). Available on 15th April 2019 at: https://www.earlymoderntexts.com/

YOUNG, I. M., *Justice and the Politics of Difference*, Princeton, Princeton University Press, 1990.

5. Documents issued, commissioned or published by public bodies

ALSTON, P., "Digital welfare states and human rights", UN Special Rapporteur on extreme poverty and human rights, report A/74/493, 11th October 2019.

ARTICLE 29 WORKING PARTY, "Advice paper on special categories of data ('sensitive data')", 20th April 2011.

ARTICLE 29 WORKING PARTY, "Opinion 03/2013 on purpose limitation", 00569/13/EN, WP 203, 2nd April 2013.

ARTICLE 29 WORKING PARTY, "Opinion 05/2014 on Anonymisation Techniques", 0829/14/EN, 10th April 2014.

ARTICLE 29 WORKING PARTY, "Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679", 17/EN, WP 248 rev.01, 4th October 2017

ARTICLE 29 WORKING PARTY, "Guidelines on automated individual decision-making and profiling for the purpose of Regulation 2016/679", 17/EN, WP 251rev.01, 6th February 2018.

ARTICLE 29 WORKING PARTY, "Guidelines on consent under Regulation 2016/679", 17/EN WP259 rev.01, 10th April 2018.

BOARD OF GOVERNORS OF THE FEDERAL RESERVE SYSTEM, OFFICE OF THE COMPTROLLER OF THE CURRENCY, "Supervisory Guidance on Model Risk Management", SR Letter 11-7, 2011.

CHOPIN, I. & GERMAIN, C., "A comparative analysis of non-discrimination law in Europe 2019", Brussels, European Commission, 2019.

COUNCIL OF EUROPE, "Human rights and modern scientific and technological developments", Recommendation 509/1968.

COUNCIL OF EUROPE, Resolution 2111 (2016) on assessing the impact of measures to improve women's political representation of the Council of Europe's Parliamentary Assembly**.**

COUNCIL OF EUROPE COMMISSIONER FOR HUMAN RIGHTS, "Unboxing Artificial Intelligence: 10 steps to protect Human Rights", Strasbourg, Council of Europe, 2019.

DATENETHIKKOMMISSION, "Gutachten der Datenethikkommission", 2019.

DATENETHIKKOMMISSION, "Opinion of the Data Ethics Commission: Executive Summary", 2019.

DG JUSTICE "Compendium of practice on non-discrimination/equality mainstreaming", European Commission, 2011. Available on 28th April 2019 at: https://publications.europa.eu/

EU AGENCY FOR FUNDAMENTAL RIGHTS, "Handbook on European non-discrimination law", Luxembourg, Publications Office of the European Union, 2018.

EU COMMISSION, "Communication from the Commission on the precautionary principle", COM/2000/0001 final, 1st February 2000.

EU COMMISSION, "Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats)", Official Journal of the EU, 13th January 2012.

EU COMMISSION, "White Paper on Artificial Intelligence – A European approach to excellence and trust", COM(2020) 65 final, 2nd February 2020.

EUROPEAN DATA PROTECTION BOARD, "Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects", 8th October 2019.

EUROPEAN DATA PROTECTION SUPERVISOR, "Towards a new digital ethics: data, dignity and technology", 2015.

EUROPEAN PARLIAMENT, "European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics", 2015/2103(INL), 2017.

GENERAL DIRECTORATE OF THE SPANISH TAX ADMINISTRATION AGENCY, Resolution, 11th January 2019, approving the general guidelines of the Annual Tax and Customs Control Plan for 2019.

HOUSE OF COMMONS SCIENCE AND TECHNOLOGY COMMITTEE, "Algorithms in decision-making", 2018.

INFORMATION COMMISSIONER'S OFFICE, "Anonymisation: managing data protection risk code of practice", 2012.

- "Guide to the General Data Protection Regulation", 2018. Available on May 17th 2019 at: https://ico.org.uk/

IQBAL, S., "Business, women and the law 2016: getting to equal (English)", Washington D.C, World Bank Group, 2015.

KOCHI, E., "How to prevent discriminatory outcomes in machine learning", World Economic Forum Global Future Council on Human Rights, 2018.

KOENE, A., CLIFTON, C., HATADA, Y., WEBB. H., PATEL, M., MACHADO, C., LAVIOLETTE, J., RICHARDSON, R. & REISMAN, D., "A governance framework for algorithmic accountability and transparency", European Parliamentary Research Service, 2019.

NADLER, J. & CICILLINE, D. N., "Investigation of competition in digital markets", Subcommittee on antitrust, commercial and administrative law of the Committee on the judiciary, October 2020.

OECD, "Entrenched social norms prevent the equal distribution of caring responsibilities between men and women", March 2018.

OECD, "Profiling tools for early identification of jobseekers who need extra support", December 2018.

TOBLER, C., "Limits and potential of the concept of indirect discrimination", Luxembourg, Publications Office of the EU, 2008.

UK GOVERNMENT OFFICE FOR SCIENCE, "Artificial intelligence: an overview for policymakers", 2016.

UK GOVERNMENT, "Guidelines for AI procurement", 8th June 2020.

US DEPARTMENT OF EDUCATION, "Enhancing teaching and learning through educational data mining and learning analysis", October 2012.

US EXECUTIVE OFFICE OF THE PRESIDENT, "Big data: seizing opportunities, preserving values", 2014.

- "Big data and differential pricing", 2015.

- "Artificial intelligence, automation and the economy", 2016.

US FEDERAL TRADE COMMISSION, "In FTC study, five percent of consumers had errors on their credit reports that could result in less favorable terms for loans", 11th February 2013. Available on 9th April 2019 at: https://www.ftc.gov/

US SENATE COMMITTEE ON COMMERCE, SCIENCE & TRANSPORTATION, MAJORITY STAFF, "A review of the data broker industry: collection, use, and sale of consumer data for marketing purposes", 18th December 2013.

US SENATE HEALTH, EDUCATION, LABOR AND PENSIONS COMMITTEE, "For profit higher education: the failure to safeguard the federal investment and ensure student success", 2012.

VICKERS, L., "Religion and belief discrimination in employment – the EU Law", Luxembourg, Office of Publications of the European Communities, 2006.

WEISS, M. A., & ARCHICK, K., "US-EU data privacy: from safe harbor to privacy shield", *Congressional Research Service*, 19th May 2016.

WORLD ECONOMIC FORUM, "Guidelines for AI Procurement", 2019.

ZUIDERVEEN BORGESIUS, F., "Discrimination, artificial intelligence and algorithmic decision-making", Strasbourg, Directorate General of Democracy, Council of Europe, 2018.

6. Private sector reports, press releases and official documents

ALPHABET, "Alphabet announces second quarter 2020 results", 2020. Available on 12th September 2020 at: https://abc.xyz/

ALLEN, M., "Health insurers are vacuuming up details about you — and it could raise your rates", *Propublica*, 17th July 2018.

ANGWIN, J., LARSON, J., MATTU, S., & KIRCHNER, L., "Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks", *Propublica*, 23rd May 2016.

ANGWIN, J. & PARRIS JR., T., "Facebook lets advertisers exclude users by race", *Propublica*, 28th October 2016.

ANGWIN, J., TOBIN, A. & VARNER, M., "Facebook (still) letting housing advertisers exclude users by race", *Propublica*, 21st November 2017.

APPLE, "Inclusion and diversity", 2017. Available on 27th April 2019 at: https://www.apple.com/diversity/

ATSKE, S., GEIGER, A. & SCHELLER, A., "The share of women in legislatures around the world is growing, but they are still underrepresented", *Pew Research Center*, 18th March 2019.

BECKETT, L., "Everything we know about what data brokers know about you", *Propublica*, 13th June 2014.

BERNACIAK, M., GUMBRELL-MCCORMICK, R., AND HYMAN, R., "European trade unionism: from crisis to renewal?", European Trade Union Institute, Report No. 133.

BERTRAND, M., MULLAINATHAN, S. & ABRAMS, D., "Discrimination in the judicial system", Innovations for Poverty Action, 2001.

BOGEN, M. & RIEKE, A., "Help wanted: an examination of hiring algorithms, equity and bias", *Upturn*.

BURRI, S., SENDEN, L. & TIMMER, A., "A comparative analysis of gender equality law in Europe 2019", Brussels, European Commission, 2019.

CONSUMER REPORTS, "The secret score behind your auto insurance", 10th August 2006.

EY, "The new age: artificial intelligence for human resource opportunities and functions", 2019. Available on 3rd April 2019 at: https://www.ey.com/

FACEBOOK, "Facebook Q2 2020 results", 2020.

- "Improving enforcement and promoting diversity: updates to ads policies and tools", *Facebook Newsroom*, 8th February 2017.

GILLUM, J. & TOBIN, A., "Facebook won't let employers, landlords or lenders discriminate in ads anymore", *Propublica*, 19th March 2019.

GOOGLE, "Google diversity annual report 2018", 2018.

GROSS, S. R., POSSLEY, M. & STEPHENS, K., "Race and wrongful convictions in the United States", National Registry of Exonerations, 7th May 2017.

HUMAN RIGHTS WATCH, "Mind the gap: the lack of accountability of killer robots", 2015.

HUNCHLAB, "A citizen's guide to HunchLab", 11th July 2017.

IDEA, "Gender quotas database", 2019. Available on 28th April 22019 at: https://www.idea.int/

INGOLD, D. & SOPER, S., "Amazon doesn't consider the race of its customers. Should it?", *Bloomberg*, 21st April 2016.

MCINTYRE, L., "Diversity and inclusion update: The journey continues", *Microsoft*, 14th November 2018.

NEW YORK CIVIL LIBERTIES UNION, "Stop-and-frisk 2011", 2012.

NORDEN, "Gender equality – The nordic way", Copenhaguen, Nordic Council of Ministers, 2010.

NORTHPOINTE, "Risk assessment". Available on 27th March 2019 at: https://www.documentcloud.org/

RICHARDSON, R. (ed.), "Confronting black boxes: a shadow report of the new york city automated decision system task force," AI Now Institute, 2019.

SULLIVAN, L., MESCHEDE, T., DIETRICH, L, SHAPIRO, T., TRAUB, A., RUETSCHLIN, C. & DRAUT, T., "The racial wealth gap: why policy matters", IASP/Demos, 2015.

ZICK, A., KÜPPER, B. & HÖVERMANN, A., "Intolerance, Prejudice and Discrimination: A European Report", Berlin, Nora Langenbacher, 2011.

7. News and magazine articles, blog posts and information published on websites

ALGORITHMIC JUSTICE LEAGUE. Available on 17th October 2019 at: https://www.ajlunited.org/

ALGORITHM REGISTER, "Housing rental fraud risk", 2020. Available on 9th October 2020 at: https://algoritmeregister.amsterdam.nl/

ALGORITHM WATCH. Available on 17th October 2019 at: https://algorithmwatch.org/en/

ARROYO JIMÉNEZ, L., "Algoritmos y reglamentos", *Almacén de Derecho*, 25th February 2020. Available on 1st April 2020 at: https://almacendederecho.org/

BBVA, "How credit scoring can influence the granting of a loan". Available on 25th February 2019 at: https://www.bbva.es/

BELMONTE, E., "La aplicación del bono social del Gobierno niega la ayuda a personas que tienen derecho a ella", *CIVIO*, 16th May 2019. Available on 6th December 2019 at: https://civio.es/

BLACK GIRLS CODE, "What we do", 2018. Available on 13th June 2019 at: http://www.blackgirlscode.com/

BHATIA, R., "How do machine learning algorithms differ from traditional algorithms?", *Analytics India Magazine*, 10th September 2018. Available on 13th June 2019 at: https://analyticsindiamag.com/

CADOT, J., "Lesbienne : Google a enfin modifié son algorithme", *Numerama*, 18th July 2019. Available on 3rd March 2020 at: https://www.numerama.com/

CAPALA, M., "Global search engine market share for 2018 in the top 15 GDP nations", 27th August 2018, *Alphametic*. Available on 10th April 2019 at: https://alphametic.com/

COASTINE, J., "Facebook is shutting down its API for giving your friends' data to apps", *TechCrunch*, 28th April 2015. Available on 13th June 2020 at: https://techcrunch.com/

CORBETT-DAVIES, S., PIERSON, E. & GOEL, S., "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear", *The Washington Post*, 17th October 2015. Available on 9th April 2019 at: https://www.washingtonpost.com/

COUNCIL OF EUROPE, "Prohibition of discrimination", 2017. Available on 5th June 2019 at: https://www.coe.int/

CRAWFORD, K., "Artificial Intelligence's White guy problem", *The New York Times*, 25th June 2016. Available on 5th April 2019 at: https://www.nytimes.com/

CAREERBUILDER, "More than half of HR managers say artificial intelligence will become a regular part of HR in next 5 years", 18th May 2017. Available on 12th February 2019 at: https://www.prnewswire.com/

CIVIO, "Que se nos regule mediante código fuente o algoritmos secretos es algo que jamás debe permitirse en un Estado social, democrático y de Derecho", *CIVIO*, 2nd July 2019. Available on 6th December 2019 at: https://civio.es/

CREDIT SCORE DATING. Available on 27th March 2019 at: www.creditscoredating.com

DASTIN, J., "Amazon scraps secret AI recruiting tool that showed bias against women", *Reuters,* 10th October 2018. Available on 12th February 2019 at: https://www.reuters.com/

DAVIES, D., "How search engine algorithms work: everything you need to know", *Search Engine Journal*, 10th May 2018. Available on 3rd April 2019 at: https://www.searchenginejournal.com/

DEL CASTILLO, C., "Contra la violencia machista, el odio y las denuncias falsas: los algoritmos que usa la Policía", *eldiario.es*, 1st January 2019. Available on 11th April 2019 at: https://www.eldiario.es/

DJEFFAL, C., "Deutschland braucht nicht ein Digitalministerium, sondern viele!", *Süddeutsche Zeitung*, 18th September 2017. Available on 21st May 2020 at: https://www.sueddeutsche.de/

DOUGHERTY, C., "Google photos mistakenly labels Black people 'gorillas'", *The New York Times BITS Blog*, 1st July 2015. Available on 11th April 2019 at: https://bits.blogs.nytimes.com/

ECKERD CONNECTS, "Eckerd rapid safety feedback". Available on 4th April 2019 at: https://eckerd.org/

EMPRESA MUNICIPAL DE LA VIVIENDA Y EL SUELO, "Procedimiento de adjudicación", *Ayuntamiento de Madrid*. Available on 27th November 2019 at: https://www.emvs.es/

EUROCOP, "Análisis y prevención del delito", 2015. Available on 11th April 2019 at: https://www.eurocop.com/

EUROPEAN COMMISSION, "Antitrust: Commission fines Google €1.49 billion for abusive practices in online advertising", 20th March 2019. Available on 20th June 2020 at: https://ec.europa.eu/

EUROPEAN DATA PROTECTION BOARD, "GDPR: Guidelines, recommendations, best practices". Available on 11th May 2019 at: https://edpb.europa.eu/

- "Hamburg commissioner fines H&M 35.3 million euro for data protection violations in service centre", 2020. Available on 3rd October 2020 at: https://edpb.europa.eu/

FELTEN, E., "What does it mean to ask for an 'explainable' algorithm?", *Freedom to Tinker*, 31st May 2017. Available on 9th July 2019 at: https://freedom-to-tinker.com/

FURLOW, B., "IBM Watson collaboration aims to improve oncology decision support tools", *The Journal of Oncology*, 16th March 2016. Available 3rd April 2019 at: https://www.cancernetwork.com/

FUTURE OF LIFE, "Asilomar AI principles", 2017. Available on 14th March 2020 at: https://futureoflife.org/

GDPR Enforcement Tracker. Available on 6th July 2020 at: https://www.enforcementtracker.com

Google Ad Settings. Available on 11th April 2019 at: https://adssettings.google.com/

GOOGLE AI, "Artificial intelligence at Google: Our principles", 2019. Available on 19th September 2019 at: https://ai.google/principles/

- "Artificial intelligence at Google: Responsible AI practices", 2019. Available on 19th September 2019 at: https://ai.google/principles/

GROSSFELD, B., "A simple way to understand machine learning vs deep learning", Zendesk, 18th July 2017. Available on 31st January 2019 at: https://www.zendesk.com/

HARDT, M., "How big data is unfair", *Medium*, 26th September 2014. Available on 6th May 2019 at: https://medium.com/

HENN, S., "When women stopped coding", *Planet Money,* 21st October 2014. Available on 26th April 2019 at: https://www.npr.org/

HUNT, B., "Redlining", *Encyclopedia of Chicago*, 2005. Available on 20th February 2019 at: http://www.encyclopedia.chicagohistory.org/

IBORDERCRTL, "Technical Framework", 2016. Available on 16th May 2019 at: https://www.iborderctrl.eu/

IEEE. Available on 2nd October 2019 at: https://www.ieee.org

- "IEEE P7003 standards for algorithmic bias considerations", *IEEE*. Available on 2nd October 2019 at: https://standards.ieee.org/

INFORMATION COMMISSIONER'S OFFICE, "Intention to fine British Airways £183.39m under GDPR for data breach", 8th July 2018. Available on 6th July 2020 at: https://ico.org.uk/

INTERNATIONAL ASSOCIATION FOR IMPACT ASSESSMENT, "IAIA: Leading the global network for impact assessment", 2019. Available on 28th October 2019 at: https://www.iaia.org

ISO, "Standards". Available on 3rd October 2019 at: https://www.iso.org/

JUNQUERAS, O., "Proximitats genètiques", *Avui*, 27th August 2008. Available on 26th April 2019: https://s.libertaddigital.com/

KAPOR CENTER, "Tech workforce". Available on 22nd January 2020 at: https://leakytechpipeline.com/

KAYNE, C., "Do credit scores matter outside the US?", *CNBC*, 9th February 2011. Available on 25th February 2019 at: https://www.cnbc.com/

KELLING, G. L. & WILSON J. Q., "Broken windows: the police and neighborhood safety", *Atlantic Monthly*, March 1982. Available on 14th February 2019 at: https://www.theatlantic.com/

KOLOTÚSHKINA, N., "VERIPOL: la herramienta de la Policía para detectar denuncias falsas", *RTVE*, 2nd November 2018. Available on 19th February 2019 at: http://www.rtve.es/

LÓPEZ ZAFRA, J. M., "Patrones de comportamiento y voracidad fiscal", *El Confidencial*, 14th July 2018. Available on 16th July 2019 at: https://blogs.elconfidencial.com/

LOHR, S., "Sizing up big data, broadening beyond the Internet", *The New York Times BITS Blog*, 29th June 2013. Available on 28th April 2019 at: https://bits.blogs.nytimes.com/

LYTVYNOVA, K., "Machine learning project structure: stages, roles, and tools". Available on 7th April 2019 at: https://datafloq.com/

MALAN, D., "What is an algorithm?", May 2013. Available on 27th August 2020 at: https://www.ted.com/

MARWICK, A. E., "How your data are being deeply mined", *New York Review of Books*, 9th January 2014. Available on 20th February 2019 at: http://www.tiara.org/

MORRISON, T., "Making America white again", *The New Yorker*, 14th November 2016. Available on 13th April 2019 at: https://www.newyorker.com/

MYDATA, "Who we are", 2018. Available on 5th May 2019 at: https://mydata.org/about/

O'NEIL RISK CONSULTING AND ALGORITHMIC AUDITING (ORCAA), "Services", 2019. Available on 19th September 2019 at: http://www.oneilrisk.com

PADOFF, R., "Why deep learning is suddenly changing your life", *Fortune*, 28th September 2016. Available on 28th April 2019 at: http://fortune.com/

PAYNE, A., "Credit score systems across the world", *Graydon*, 9th February 2015. Available on 25th February 2019 at: https://www.graydon.co.uk/

PECK, D., "They're watching you at work", *The Atlantic*, December 2013. Available on 20th February 2019 at: https://www.theatlantic.com/

PINET, J. P. & REDEGELD, T., "Poverty discrimination in Europe", 2018. Available on 28th March 2020 at: https://blogs.atd-quartmonde.org/

PLANET LABOR, "Austria: an algorithm that evaluates the unemployed (briefly)", 24th October 2018. Available on 23rd January 2020 at: https://www.planetlabor.com/

PREDPOL. Available, on February 14th 2019 at: https://www.predpol.com

REDACCIÓN MÉDICA, "'Big data' e IA mejoran un 40% la detección precoz de la sepsis grave", 10th March 2019. Available on 3rd April 2019 at: https://www.redaccionmedica.com/

SEOMARK, "How does Google rank websites?", *SEOmark*, 20th September 2019. Available on 11th April 2019 at: https://www.seomark.co.uk/

SHEAD, S., "Google DeepMind is giving the NHS free access to its patient monitoring app", *Business Insider*, 24th June 2017. Available 5th on April 2019 at: https://www.businessinsider.de/

SIMONITE, T., "When it comes to gorillas, Google photos remains blind", *Wired*, 11th January 2018. Available on 11th April 2019 at: https://www.wired.com/

STOPFAKES.GOV (INTELLECTUAL PROPERTY RIGHTS INFORMATION & ASSISTANCE), "Why is intellectual property important?", 7th July 2016. Available on 22nd May 2019 at: https://www.stopfakes.gov/

SZIGETVARI, A., "Arbeitsmarktservice gibt grünes Licht für Algorithmus", *Der Standard*, September 17th, 2019. Available on January 23rd 2020 at: https://www.derstandard.at/

TAYLOR, C., "Structured vs. Unstructured data", *Datamation*, 28th March 2018. Available on 13th June 2019 at: https://www.datamation.com/

TEICH, P., "Artificial intelligence can reinforce bias, cloud giants announce tools for AI fairness", *Forbes*, 24th September 2018. Available on 19th September 2019: https://www.forbes.com/

TELFORD, T., "Apple Card algorithm sparks gender bias allegations against Goldman Sachs", 11th November 2019. Available on 23rd January 2020 at: https://www.washingtonpost.com/

THOMAS, E., "Why Oakland police turned down predictive policing", *VICE*, 28th December 2016. Available on 2nd April 2020 at: https://www.vice.com/

UN WOMEN, "UN Women ad series reveals widespread sexism", 21st October 2013. Available on 10th April 2019 at: http://www.unwomen.org/

VALENTINO-DEVRIES, J., SIGER-VINE, J. & SOLTANI, A., "Websites vary prices, deals based on users' information", *The Wall Street Journal*, 24th December 2012. Available on 12th February 2019 at: https://www.wsj.com/

VIÑAS COLL, J., "Así son los superordenadores de Montoro contra el fraude fiscal", *Cinco Días*, 24th July 2015. Available on 16th July 2019 at: https://cincodias.elpais.com/

WILLIAMS, M., "Facebook 2018 diversity report: reflecting on our journey", *Facebook Newsroom*, 12th July 2018. Available on 27th April 2019 at: https://newsroom.fb.com/

WORLD INTELLECTUAL PROPERTY ORGANISATION, "What is intellectual property?", 2004. Available on 22nd May 2019 at: https://www.wipo.int/

ZARYA, V., "Why being a woman hurts your credit score", *Fortune*, 10th February 2016. Available on 19th April 2019 at: http://fortune.com/

8. Case law

Administrative regional court of Lazio-Roma, Section III bis, Judgment No. 3769, 22nd March 2017.

Administrative regional court of Lazio-Roma, Section III bis, Judgment No. 10964, 13th September 2019.

District Court of The Hague, ruling of February 5th 2020, case number C / 09/550982 / HA ZA 18-388.

Court of 1st Instance of Namur, May 5th 2015, Le centre interfédéral pour l'égalité des chances et la lutte contre le racisme et les discriminations v. M. Cristophe.

ECHR Judgment 21st February 1990, 9310/81, Powell and Rayner v. The United Kingdom.

ECHR Judgment 9th December 1994, 16798/90, López Ostra v. Spain.

ECHR Judgment 12th June 2003, 35968/97, Van Kück v. Germany.

ECHR Judgment 13th December 2005 (final decision March 13th 2006), 55762/00 and 55974/00, Timishev v. Russia.

ECHR Judgment 10th May 2007, 42949/98 and 53134/99, Runkee and White v. United Kingdom.

ECHR Judgment 13th November 2007, 57325/00, D.H. and others v. the Czech Republic.

ECHR Judgment 17th February 2011, 6268/08, Andrle v. Czech Republic.

ECHR Judgment 24th July 2012, 47159/08, B.S. v. Spain, 47159/08.

ECHR Judgment 11th March 2014, 26827/08, Abdu v. Bulgaria.

ECHR Judgment 30th June 2016, 51362/09, Taddeucci and Mccall v. Italy.

ECHR Judgment 24th January 2017, 60367/08, Khamtokhu and Aksenchik v. Russia.

ECHR Judgment, 25th July 2017, 17484/15, Judgment Carvalho Pinta v. Portugal.

CJEU Judgment 8th April 1976, C-43/75, Gabrielle Defrenne v. Société anonyme belge de navigation aérienne Sabena.

CJEU Judgment 13th May 1986, C-170/84, Bilka – Kaufhaus GmbH v. Karin Weber von Hartz.

CJEU Judgment 21st May 1985, C-248/83, Commission of the European Communities v. Federal Republic of Germany.

CJEU Judgment 15th May 1986, C-222/84, Johnston v. Chief Constable of the Royal Ulster Constabulary.

CJEU Judgment 13th July 1989, C-215/88, Casa Fleischhandel v. Bundesantalt für landwirtschaftliche marktordnung.

CJEU Judgment, 17th October 1989, C-109/88, Union of Commercial and Clerical Employees, Denmark v. Danfoss A/S.

CJEU Judgment 8th November 1990, C-177/88, Elisabeth Johanna Pacifica Dekker v. Stichting Vormingscentrum voor Jong Volwassenen (VJV Centrum) Plus.

CJEU Judgment 31st May 1995, C-400/93, Specialarbejderforbundet i Danmark v. Dansk Industri, formerly Industriens Arbejdsgivere, acting for Royal Copenhagen A/S.

CJEU Judgment 17th October 1995, C-450/93, Eckhard Kalanke v. Freie Hansestadt Bremen.

CJEU Judgment, 11th November 1997, Case C-409/95, Hellmut Marschall v. Land Nordrhein Westfalen.

CJEU Judgment 9th February 1999, C-167/97, Regina v. Secretary of State for Employment, *ex parte* Nicole Seymour-Smith and Laura Perez.

CJEU Judgment 11th May 1999, C-309/97, Angestelltenbetriebsrat der Wiener Gebietskrankenkasse v. Wiener Gebietskrankenkasse.

CJEU Judgment 3rd February 2000, C-207/98, Mahlburg v. Land Mecklenburg-Vorpommern.

CJEU Judgment 6th July 2000, C-407-98, Abrahamsson v. Fogelqvist.

CJEU Judgment 28th March 2000, C-158/97, Badeck v. Landesanwalt beim Staatsgerichtshof des Landes Hessen.

CJEU Judgment 26th June 2001, C-381/99, Susanna Brunnhofer v. Bank der österreichischen Postsparkasse AG.

CJEU Judgment 19th March 2002, C-476/99, H. Lommers v. Minister van Landbouw, Natuurbeheer en Visserij.

CJEU Judgment 13th January 2004, C-256/01, Debra Allonby v. Accrington & Rossendale College, Education Lecturing Services, trading as Protocol Professional and Secretary of State for Education and Employment.

CJEU Judgment April 1st 2008, C-267/06, Tadao Maruko v. Versorgungsanstalt der deutschen Bühnen.

CJEU Judgment 10th July 2008, C-54/07, Firma Feryn NV v. Centrum voor gelijkheid van kansen en voor racismebestrijding.

CJEU Judgment 17th July 2008, C-303/06, S. Coleman v. Attridge Law and Steve Law.

CJEU Judgment 16th December 2008, C-524/06, Heinz Huber v. Bundesrepublik Deutschland.

CJEU Judgment 12th January 2010, C-229/08, Colin Wolf v. Stadt Frankfurt am Main.

CJEU Judgment 12th January 2010, C-341/08, Domnica Petersen v. Berufungsausschuss für Zahnärzte für den Bezirk Westfalen-Lippe.

CJEU Judgment 1st March 2011, C-236/09, Association belge des Consommateurs Test-Achats ASBL, Yann van Vugt, Charles Basselier v. Conseil des ministres.

CJEU Judgment 21st July 2011, C-159/10 and C-160/10, Gerhard Fuchs and Peter Köhler v. Land Hessen.

CJEU 13th September 2011, C-447/09, Reinhard Prigge and Others v. Deutsche Lufthansa AG.

CJEU Judgment April 19th 2012, C-415/10, Galina Meister v. Speech Design Carrier Systems GmbH.

CJEU Judgment 22nd November 2012, C-385/11, Isabel Elbal Moreno v. Instituto Nacional de la Seguridad Social, Tesorería General de la Seguridad Social.

CJEU Judgment 28th February 2013, C-427/11, Margaret Kenny and others v. Minister for Justice, Equality and Law Reform, Minister for Finance, Commissioner of An Garda Síochána.

CJEU Judgment 25th April 2013, C-81/12, Asociaţia Accept v. Consiliul Naţional pentru Combaterea Discriminării.

CJEU Judgment 18th March 2014, C-167/12, C.D. v. S.T.

CJEU Judgment 8th April 2014, Joined Cases C-293/12 and C-594/12, Digital Rights Ireland Ltd v. Minister for Communications, Marine and Natural Resources and Others and Kärntner Landesregierung and Others.

CJEU Judgment 13th May 2014, C- 131/12 Google Spain SL and Google Inc. v. Agencia Española de Protección de Datos (AEPD) and Mario Costeja González.

CJEU Judgment 17th July 2014, C-173/13, Maurice Leone, Blandine Leone v. Garde des Sceaux, ministre de la Justice, Caisse nationale de retraite des agents des collectivités locales.

CJEU Judgment 11th November 2014, C-530/13, Leopold Schmitzer v. Bundesministerin für Inneres.

CJEU Judgment 14th April 2015, C-527/13, Lourdes Cachaldora Fernández v. Instituto Nacional de la Seguridad Social.

CJEU Judgment 16th July 2015, C-83/14, CHEZ Razpredelenie Bulgaria AD v. Komisia za zashtita ot diskriminatsia,

CJEU Judgment 16th July 2015, C-222/14, Konstantinos Maïstrellis v. Ypourgos Dikaiosynis, Diafaneias kai Anthropinon Dikaiomaton.

CJEU Judgment 6th October 2015, C-362/14, Maximillian Schrems v. Data Protection Commissioner and Digital Rights Ireland Ltd.

CJEU Judgment 24th November 2016, C-443/15, David L. Parris v. Trinity College Dublin and Others.

CJEU Judgment 14th March 2017, C-157/15, Samira Achbita and Centrum voor gelijkheid van kansen en voor racismebestrijding v. G4S Secure Solutions NV.

CJEU Judgment 14th March 2017, C-188/15, Asma Bougnaoui and Association de défense des droits de l'homme (ADDH) v. Micropole SA.

CJEU Judgment, 17th April 2018, C-414/16, Vera Egenberger v. Evangelisches Werk für Diakonie und Entwicklung e.V.

CJEU Judgment 24th September 2019, C-136/17, GC and Others v. Commission nationale de l'informatique et des libertés (CNIL).

CJEU Judgment 24th September 2019, C-507/17, Google LLC, successor in law to Google Inc. v. Commission nationale de l'informatique et des libertés (CNIL).

CJEU Judgment 16th July 2020, C-311/18, Data Protection Commissioner v. Facebook Ireland Ltd. and Maximillian Schrems.

Court of First Instance (Third Chamber) Judgment 11th September 2002, C-T-13/99, Pfizer Animal Health SA v. Council of the European Union, paragraph 139.

Spanish Constitutional Court Judgment No. 292/2000, 30th November.

Spanish Constitutional Court Judgment No. 145/1991, 1st July.

US District Court for the Northern District of Georgia, Atlanta division, "Complaint for permanent injunction and other equitable relief at 35 FTC v. Compucredit Corp", No. 1:08-CV-1976-BBM, 2008.

US Supreme Court, *Griggs v. Duke Power Co*., 401 U.S. 424, 1971.

9. Administrative bodies' resolutions and EU advocate generals' opinions

Catalan Commission for the Guarantee of the right of access to public information, Joined decisions 123/2016 and 124/2016.

European Committee of Social Rights Decision 11th September 2013, Complaint No. 81/2012, European Action of the Disabled (AEH) v. France.

French Commission on access to administrative documents, decisions No. 20144578 of 8th January 2015 and No. 20180276 of 19th April 2018.

Spanish Transparency and Good Government Council, Resolution 701/2018, February 18th.

Opinion of Advocate General Léger delivered on 31st May 1995 on Case C-317/93 Inge Nolte v Landesversicherungsanstalt Hannover.

Opinion of Advocate General Kokott delivered on 30th September 2010 on Case C-236/09 Association belge des Consommateurs Test-Achats ASBL, Yann van Vugt, Charles Basselier v. Conseil des ministres.

Opinion of Advocate General Kokott delivered on 31st May 2016 on Case C-157/15 Samira Achbita and Centrum voor gelijkheid van kansen en voor racismebestrijding v. G4S Secure Solutions NV.

9. Rules and regulations

9.1. Council of Europe

Convention 108 for the protection of individuals with regard to the processing of personal data.

European Convention on Human Rights.

Protocol 12 to the European Convention on Human Rights.

9.2. European Union

Directive 96/61/EC of 24 September 1996 concerning integrated pollution prevention and control.

Council Directive 2000/43/EC implementing the principle of equal treatment between persons irrespective of racial or ethnic origin.

Council Directive 2000/78/EC against discrimination at work on grounds of religion or belief, disability, age or sexual orientation.

Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to environmental information and repealing Council Directive 90/313/EEC.

Directive 2003/35/EC of the European Parliament and of the Council of 26 May 2003 providing for public participation in respect of the drawing up of certain plans and programmes relating to the environment and amending with regard to public participation and access to justice Council Directives 85/337/EEC and 96/61/EC - Statement by the Commission.

Council Directive 2004/113/EC of 13 December 2004 on equal treatment for men and women in the access to and supply of goods and services.

Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast).

Directive 2006/114/EC of the European Parliament and of the Council of 12 December 2006 concerning misleading and comparative advertising

Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs.

Directive 2010/41/EU of the European Parliament and of the Council of 7 July 2010 on the application of the principle of equal treatment between men and women engaged in an activity in a self-employed capacity and repealing Council Directive 86/613/EEC.

Directive 2010/75/EU of 24 November 2010 on industrial emissions (integrated pollution prevention and control).

Directive 2012/13/EU of the European Parliament and of the Council of 22 may 2012 on the right to information in criminal proceedings.

Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA.

Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure.

Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law.

Decision (EU) 2016/1250 of 12 July 2016 pursuant to Directive 95/46/EC of the European Parliament and of the Council on the adequacy of the protection provided by the EU-U.S. Privacy Shield (notified under document C(2016) 4176).

Charter of Fundamental Rights of the European Union.

Regulation (EU) 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation, amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC.

9.3.Spain

Act 19/2013, 9th November, on transparency, access to public information and good government.

Act 22/2018 of the Valencian government on the general inspection of services and on the system of alerts for the prevention of bad practices in the Valencian public Administration and its instrumental public sector.

Organic Act 2/2018 modifying the General Electoral Regime Organic Act 5/1985 to guarantee the right of suffrage of all persons with disabilities.

Organic Act 3/2007 for the effective equality of women and men.

Regulation 20th December 2018, for the adjudication of housing managed by the municipal housing and land company of Madrid.

Royal Decree 897/2017, 6th October, which regulates the concept of vulnerable consumer, the social bond and other protection measures for domestic electricity consumers.

Royal Legislative Decree 1/1996, 12th April 1996, which passes the revised text of the Law on Intellectual Property, regularising, clarifying and harmonising the legal provisions in force on the subject.

9.4.Other national regulatory instruments

Bulgarian Protection Against Discrimination Act.

Canadian Directive on Automated Decision-Making, 1st April 2019.

Croatian Anti-discrimination Act.

French Act No. 2016-832, 24th June 2016, aimed towards fighting discrimination on the basis of social precariousness.

German Federal Data Protection Act of 30th June 2017.

German General Act on Equal Treatment.

UK Data Protection Act 2018.