Revista de
Estudios
Kantianos

Revista de
Estudios
Kantianos

Revista de Estudios Kantianos

Publicación internacional de la Sociedad de Estudios Kantianos en Lengua Española
Internationale Zeitschrift der Gesellschaft für Kant-Studien in Spanischer Sprache
International Journal of the Society of Kantian Studies in the Spanish Language

Entidades colaboradoras
Sociedad de Estudios Kantianos en Lengua Española (SEKLE)
Departament de Filosofia de la Universitat de València
Instituto de Humanidades, Universidad Diego Portales

Índice

Artículos

Semblanza

Recensiones

Artículos

¿Puede la razón práctica ser artificial?[1]

DIETER SCHÖNECKER[2]

Resumen

¿Puede la razón práctica ser artificial? Desde un punto de vista kantiano, la respuesta es claramente negativa: la razón práctica no puede ser artificial. Luego de una observación preliminar acerca de la posibilidad de las máquinas morales kantianas (1.1) y de ciertos aspectos fundamentales acerca del concepto de razón práctica (1.2) e intuicionismo en Kant (1.3), sostendré que, en un modelo kantiano de obligación moral, el sujeto moral (humano) típico tiene, y debe tener, sentimientos morales para poder conocer la validez de la ley moral como un imperativo categórico (1.4). Por medio del argumento del conocimiento en contra del fisicalismo y del funcionalismo, sostendré que las computadoras no tienen sentimientos y, *a fortiori*, no tienen sentimientos morales; por lo tanto, las computadoras no son sujetos morales (1.5). Esta conclusión se basa en un *yo siento* kantiano más que en un *yo pienso* (1.6). A continuación, me ocuparé de dos problemas de ese argumento (2). Concluiré con una analogía (3): así como los planetas no vuelan, las computadoras no sienten.

Palabras clave: inteligencia artificial, razón práctica, sentimiento moral, *yo pienso* vs. *yo siento*, argumento del conocimiento

Can practical reason be artificial?

Resumen

Can practical reason be artificial? The answer, from a Kantian point of view, is clearly negative: Practical reason cannot be artificial. After a preliminary remark on the possibility of Kantian moral machines (1.1) and some basics on the concept of practical reason (1.2) and Kant's intuitionism (1.3), I will argue that in a Kantian model of moral obligation, the typical (human) moral subject has moral feelings and must have them in order to cognize the validity of the moral law as a categorical imperative (1.3). Using the knowledge argument against physicalism and functionalism, I shall argue that computers have no feelings and, *a fortiori*, no moral feelings; therefore, computers are no moral subjects (1.4). This conclusion is based on a Kantian *I feel* rather than *I think* (1.5). I will then tackle two problems with this argument (2). I will conclude with an analogy (3): Just as planets do not fly, computers do not feel.

Keywords: artificial intelligence, practical reason, moral feelings, *I think* vs. *I feel*, knowledge argument

La inteligencia artificial (IA) ha generado una variedad de preguntas morales, jurídicas, económicas, políticas —en suma, preguntas *prácticas*— que precisan de una respuesta pronto: desde cómo lidiar con autos que se conducen solos hasta la IA como el fin de la raza humana debido a algún tipo de rareza tecnológica. Para solo nombrar un incidente reciente, recordemos la disputa que atañe al *Korea Advanced Institute of Science and Technology*, que, en una carta abierta, fue acusado por 50 académicos de participar en un programa para desarrollar las así llamadas "máquinas asesinas" (cf.

[1] Traducción: Fiorella Tomassini (UBA-CONICET). Una versión en inglés de este artículo será publicada en *Journal of Artificial Intelligence Humanities*.

[2] Universidad de Siegen. Contacto: dieter.schoenecker@uni-siegen.de

Wakefield 2018).[3] Dado que la IA, al menos desde estándares filosóficos, es un fenómeno bastante reciente, tanto estas preguntas prácticas o problemas así como sus respuestas posibles son bastante novedosas. Nótese, sin embargo, que estas respuestas, a su vez, dependerán de fundamentaciones que lejos están de ser vírgenes o inocentes. En ética aplicada y filosofía política, rápidamente llegamos a preguntas tradicionales y a posiciones que tenemos que discutir tanto a nivel metaético como normativo para ofrecer respuestas plausibles. Por lo tanto, no es sorprendente que en artículos sobre máquinas morales se expongan, por ejemplo, argumentos sobre el viejo y querido utilitarismo (v., e.g., Bonnefon; Shariff; Rahwan 2015).

Sin embargo, aquí las preguntas prácticas o de ética aplicada sobre la IA no son de mi interés. Más bien, la pregunta que trataré aquí pertenece esencialmente a la filosofía de la mente: ¿puede la razón práctica ser artificial? La razón práctica es entendida de la mejor manera, sostengo, como un poder genuino de conocer y querer el bien. Desde un punto de vista kantiano, la respuesta a esa pregunta es claramente negativa: *la razón práctica no puede ser artificial*. Es tentador pensar que esta respuesta tiene su fundamento en el pensamiento *epistemológico* de Kant —o bien, como Kant diría, en el pensamiento *teórico* de que la razón siempre es la razón de *alguien*—, como por ejemplo que no hay pensar sin *alguien* que piense o siempre pueda pensar *yo pienso*. Examinaré esto brevemente, pero me centraré en la filosofía práctica de Kant. También desde este punto de vista práctico, la conclusión de que la razón práctica no puede ser artificial es rápida, sólida e inevitable, pues la razón práctica es libre y las computadoras no lo son. Sin embargo, mi enfoque es diferente: se basa en la idea de que la razón moral está acompañada de sentimientos morales que las computadoras no pueden tener. Luego de una observación preliminar acerca de la posibilidad de las máquinas morales kantianas (1.1) y de ciertos aspectos fundamentales acerca del concepto de razón práctica (1.2) e intuicionismo en Kant (1.3), sostendré que, en un modelo kantiano de obligación moral, el sujeto moral (humano) típico tiene, y debe tener, sentimientos morales para poder conocer la validez de la ley moral como un imperativo categórico (1.4). Por medio del argumento del conocimiento en contra del fisicalismo y del funcionalismo, sostendré que las computadoras no tienen sentimientos y, *a fortiori*, no tienen sentimientos morales; por lo tanto, las computadoras no son sujetos morales (1.5). Esta conclusión se basa en un *yo siento* kantiano más que en un *yo pienso* (1.6). A continuación, me ocuparé de dos problemas de ese argumento (2). Concluiré con una analogía (3): así como los planetas no vuelan, las computadoras no sienten.

1. El argumento a partir de los sentimientos morales

Incluso sabiendo solo un poco sobre la filosofía de Kant, es fácil entender un argumento que, si es cierto, descarta claramente la posibilidad de que las computadoras tengan razón práctica. Este es el *argumento a partir de la libertad práctica trascendental*. Un bosquejo rápido sería el siguiente: la obligación moral presupone la libertad práctica trascendental de la razón práctica. Esta libertad es, hablando negativamente, la independencia de la causalidad natural o determinismo físico. Las computadoras, sin embargo, están determinadas por leyes de la física. Por lo tanto, no pueden ser libres. Pero la razón práctica —y, por consiguiente, el ser humano— es libre, y precisa ser libre para que la moralidad tenga sentido (Kant es un incompatibilista).[4] Por lo tanto, ninguna computadora puede tener razón práctica. Nótese que incluso en una comprensión no-determinista de la física, e incluso en relación con computadoras cuánticas, este argumento a partir de la libertad sigue valiendo.

[3] Quisiera agradecer a los organizadores de *The First International Conference on Artificial Intelligence Humanities* que tuvo lugar en la Universidad Chung-Ang (Seúl), el 16 de agosto de 2018. Agradezco especialmente al Prof. Chan Kyu Lee y al Prof. Hyeongjoo Kim. Quisiera también expresar mi gratitud a Larissa Berger, al Prof. Dr. Markus Lohrey, a Christian Prust, Elke Schmidt y al Dr. Thomas Sukopp por las discusiones sobre IA.

[4] Kant es un compatibilista solo en el sentido de que la libertad y el determinismo son compatibles, asumiendo la diferencia entre el mundo nouménico y el mundo sensible. Pero esta no es la suposición común utilizada en el debate acerca del compatibilismo e incompatibilismo (cf. Schönecker 2005).

Pues, la libertad no es solo independencia de causas naturales sino además una facultad positiva, a saber, la facultad de determinarse a uno mismo en un acto autónomo de espontaneidad absoluta, y tal espontaneidad, a diferencia del azar, no es carente de ley.

Por consiguiente, solo con saber un poco de Kant, se puede ver claramente que, al menos desde el punto de vista kantiano, es bastante obvio que una computadora no puede tener razón práctica. Dado que esto es tan obvio, quisiera llamar la atención sobre otro argumento de la filosofía práctica de Kant que usualmente pasa desapercibido; lo llamaré *el argumento a partir de los sentimientos morales*.

1.1. Una observación premilimar: las máquinas morales kantianas

Una vez, Alan Turing listó un número de cosas que las personas creen que las computadoras nunca serán capaces de hacer; entre ellas, estaba la capacidad de "distinguir el bien del mal". Por supuesto, depende de qué significa "distinguir el bien del mal", pero al menos en relación con el *output* de ese "*distinguir* el bien del mal", esa presuposición puede resultar tranquilamente falsa. Desde hace algún tiempo, existen discusiones serias sobre las "máquinas morales" (cf., e.g., Wallach; Allen 2009) y el desarrollo de robots genera preguntas morales que no solo son de interés teórico (o filosófico), por así decirlo, sino que esas preguntas son tratadas de modo bastante práctico. Los algoritmos morales parecen posibles y esos algoritmos no solo, por ejemplo, pueden ayudar a los jueces a tomar decisiones morales, sino que además, en algún sentido muy limitado, pronto tomarán decisiones morales por sí mismos (pensemos tan solo en los así llamados vehículos autónomos). Tal vez sea muy tentador asumir que esas máquinas morales deben estar basadas en algún tipo de razonamiento utilitarista, dado el carácter matemático (y *prima facie* la facilidad) del cálculo utilitario o hedonista. Sin embargo, dado el aspecto *formal* del famoso imperativo categórico de Kant y de la idea de universalización, eso podría ser un prejuicio: una computadora también podría ser capaz de realizar un algoritmo moral sobre fundamentos kantianos. Recordemos la idea básica de la así llamada fórmula de la ley natural: supongamos que alguien tiene una máxima, por ejemplo, de que se suicidará si su vida causa irremediablemente, o causará, más sufrimiento que placer. El imperativo categórico obliga a que se pregunte si esa máxima puede ser una ley universal de modo tal que todo aquel que experimente más sufrimiento que placer realmente se suicide. Entonces puede que se dé cuenta de que ello lleva a algún tipo de contradicción. Hubo un largo debate, que aún continúa, sobre cómo entender la contradicción que Kant tiene en mente pero, al menos, en una interpretación de algún modo formal (lógica) de la contradicción envuelta, parece posible una máquina moral kantiana que corra un test de universalización.[5]

1.2. El concepto de razón práctica en Kant[6]

La "razón práctica" es volición buena (pura): «cada cosa de la naturaleza actúa según leyes. Sólo un ser racional tiene la facultad de obrar *según la representación* de las leyes, es decir, según principios, o [lo que es lo mismo, tiene] una *voluntad*. Como para la derivación de las acciones a partir de las leyes se requiere *razón*, entonces la voluntad no es otra cosa que razón práctica» (GMS: 412).[7] Es importante, sin embargo, diferenciar tres aspectos del concepto de razón práctica o buena voluntad en Kant: la buena voluntad de modo nouménico, la buena voluntad de modo práctico y la voluntad divina. La *buena voluntad de modo nouménico* es la voluntad autónoma que *como tal* quiere el bien. Como facultad moral, da la ley moral (el imperativo categórico) para seres imperfectos y, por medio del sentimiento moral, es también una fuerza de motivación. Todos los seres humanos tienen esta

[5] Recordemos el "felicific calculus" de Bentham.
[6] Aquí sigo a Schmidt y Schönecker (2017).
[7] Cf. GMS: 427: «la voluntad es pensada como una facultad de determinarse a sí misma,[7] a obrar *en conformidad con la representación de ciertas leyes*». Las citas en español de la *Fundamentación para la metafísica de las costumbres* son tomadas de la traducción de Fernando Moledo (Buenos Aires, Colihue, 2018, en prensa) [Nota de la traductora]. Todas las páginas y número de líneas refieren a la paginación de la Edición Académica.

ISSN-e: 2445-0669
DOI 10.7203/REK.3.2.13208

voluntad, incluso si se actúa de modo inmoral (cf. GMS: 400, 34-37; 412, 30-35; 440, 7-13; 449, 16-23; 455, 7-9.

La buena voluntad de modo nouménico es la base tanto para la buena voluntad de modo práctico como para la voluntad divina. La *buena voluntad de modo práctico* es la voluntad que tienen los seres finitos cuando sus voliciones son de hecho morales, es la buena voluntad de modo nouménico considerada como una voluntad que se manifiesta exitosamente en un ser finito contra la influencia de las inclinaciones y los deseos. Para seres imperfectos, actuar moralmente (actuar con una buena voluntad de modo práctico) significa actuar por deber. La buena voluntad de modo nouménico que se manifiesta en una persona sin obstáculos sensoriales (activos) es lo que Kant llama "la voluntad divina", solo pertenece a Dios y a otros seres divinos. Estos seres no tienen inclinaciones y deseos contrarios al bien; «la voluntad cuyas máximas concuerdan necesariamente con las leyes de la autonomía es una voluntad *santa*, absolutamente buena» (GMS: 439, 28). La buena voluntad de modo nouménico como tal (dejando de lado su incorporación a un ser finito) no puede ser diferenciada de la voluntad divina (dejando de lado su incorporación a un ser infinito). Es una causalidad nouménica: «el ser racional se incluye a sí mismo, como inteligencia, en el mundo inteligible, y llama *voluntad* a su causalidad, *meramente como una causalidad eficiente* que pertenece a ese mundo inteligible» (GMS: 453,17, el énfasis es mío). Esta voluntad se identifica luego con la voluntad que es autónoma, i.e. con la autonomía misma, «si nos pensamos como libres *nos trasladamos, como miembros, al mundo inteligible*, y reconocemos la *autonomía* de la voluntad, con su *consecuencia*, la moralidad» (GMS: 453,11, el énfasis es mío). Nótese cómo continúa Kant: «si nos pensamos como *obligados*, entonces nos consideramos como pertenecientes al mundo sensible y, sin embargo, al mismo tiempo, al mundo inteligible» (GMS: 453,14, el énfasis es mío). Por lo tanto, la voluntad libre es la voluntad nouménica (la razón práctica pura) y la autonomía es su propiedad, y en algunos contextos, esta voluntad es *considerada* no como la voluntad de un ser humano, que *también* es parte del mundo sensible, sino solo como una voluntad nouménica: «como *mero* miembro del mundo inteligible todas mis acciones serían entonces acciones perfectamente conformes al principio de la *autonomía* de la voluntad pura» (GMS: 453, 25, el énfasis es mío). Es importante tener presente que, *como tal*, la buena voluntad de modo nouménico *no* es solo una mera capacidad de actuar moralmente, pues esta voluntad como tal quiere el bien. No obstante, es la buena voluntad de modo nouménico la que *capacita* al ser *humano* para actuar moralmente. Por lo tanto, *para el ser humano* —que es un miembro del mundo nouménico *y* del mundo sensible— la buena voluntad de modo nouménico es, en efecto, una capacidad. Además, a menos que la autonomía no sea lo mismo que una buena voluntad de modo práctico, un sinvergüenza no sería autónomo —que ciertamente lo es en la medida en que incluso él, de algún modo, quiere ser moralmente bueno, i.e. en la medida en que tiene una buena voluntad *de modo nouménico*. Retomaremos este punto una vez más.

1.3. El intuicionismo en Kant

Probablemente, pronto habrá algoritmos morales por medio de los cuales las computadoras calculen qué hacer en determinadas situaciones moralmente problemáticas, pero ello no significa que *actúen* en algún sentido sustancial. Muchas veces es sorprendente ver cómo a los defensores de una IA fuerte[8] les parece obvio que las computadoras «pueden *hacer* muchas cosas igual o mejor que los seres humanos» (Russell; Norvig 2016: 1022, el énfasis es mío). Pero está claro que tal suposición comete una falacia, dado que realmente la pregunta aquí es si las computadoras pueden *hacer cualquier cosa* que un ser humano puede hacer cuando se trata de actuar o pensar. De acuerdo con Kant, no hay aquí un "hacer" en sentido estricto. Las acciones humanas, en rigor, no solo son acciones libres. Si están guiadas por la ley moral, están inmersas en sentimientos morales. Las

[8] Por "IA fuerte" me refiero, a los efectos de mis propósitos, a que una computadora podría tener conciencia, una vida interior [*qualia*] y que realmente piensa del modo en el que nosotros lo hacemos. Una computadora tal no solo imitaría el pensar, y no solo imitaría el pensar moral, sino que realmente pensaría y, por lo tanto, *pensaría moralmente*.

ISSN-e: 2445-0669
DOI 10.7203/REK.3.2.13208

computadoras no tienen sentimientos, por lo tanto, no actúan moralmente incluso si toman decisiones de acuerdo con el deber. Consideremos con más detalle esta idea.

Hasta ahora, se considera que Kant defiende, en palabras de Edmund Husserl, «un racionalismo extremo y casi absurdo», un «intelectualismo extremo» que no deja lugar a los sentimientos (1988: 412). Ahora bien, cualquier principiante en una clase introductoria de la ética de Kant aprenderá que él sostiene de manera consistente que la razón solo puede causar acciones por medio de los sentimientos. Los sentimientos, entonces, entran en juego necesariamente ya como fundamentos de determinación (motivación). Sin embargo, es importante notar que, para Kant, los sentimientos cumplen una función mucho más importante. Como hemos visto, Kant traza una línea muy estricta entre los seres divinos y los no divinos. Los seres divinos siempre quieren lo que la buena voluntad quiere, pero esto no es cierto para los seres no divinos, sensorio-racionales. Para ellos, la ley moral es siempre un imperativo categórico que los constriñe. Cito a Kant aquí con más detalle:

> Si la razón determina la voluntad indefectiblemente, entonces las acciones de un ser tal, que son reconocidas como objetivamente necesarias, son también necesarias subjetivamente; es decir, la voluntad [de un ser tal] es una facultad de elegir *sólo aquello* que la razón, independientemente de la inclinación, reconoce como prácticamente necesario, esto es, como bueno. Pero si la razón, sólo por sí misma, no determina la voluntad de manera suficiente, si la voluntad está sometida también a condiciones subjetivas (ciertos resortes impulsores) que no siempre concuerdan con las condiciones objetivas, en una palabra: si la voluntad no es *en sí* completamente conforme a la razón (que es lo que efectivamente ocurre en los hombres), entonces las acciones que objetivamente son reconocidas como necesarias, subjetivamente son contingentes, y la determinación de una voluntad tal, conforme a leyes objetivas, es *constricción*; es decir, la relación de las leyes objetivas con una voluntad que no es enteramente buena, es representada como la determinación de la voluntad de un ser racional, bien que por fundamentos de la razón, pero a los cuales esa voluntad, según su naturaleza, no es necesariamente obediente. La representación de un principio objetivo, en tanto que es *constrictivo* para una voluntad, se llama un mandamiento (de la razón) y la fórmula del mandamiento se llama *imperativo*. Todos los imperativos son expresados por medio de un *deber ser* e indican mediante él la relación de una ley objetiva de la razón con una voluntad que, según su cualidad subjetiva, no es determinada necesariamente por esa ley (una constricción) (GMS: 412ss).

El paso crucial es ver que la constricción es experimentada por el sentimiento de respeto (que, a su vez, tiene un aspecto negativo y positivo que aquí no puedo tratar, cf. Schadow 2012; Schmidt 2013). Pero este sentimiento no es solamente una suerte de efecto colateral. Dado que por "constricción" Kant se refiere a ninguna otra cosa más que al hecho de que para seres no divinos, sensorio-racionales (como nosotros), la ley es un imperativo, esto es, un deber. La obligación que tiene lugar aquí es experimentada en el sentimiento de respeto. De hecho, la obligación no solo es algo, de algún modo, *experimentado* sino además *conocido* por este sentimiento: «aquello que reconozco inmediatamente como ley para mí, lo reconozco con respeto» (GMS: 402, nota al pie). Y es importante ver que la famosa teoría de Kant del "factum de la razón" está directamente relacionada con esta idea (cf. Schönecker 2013ab). En el § 7 de la *Crítica de la razón práctica*, Kant formula el imperativo categórico. Un poco más adelante, Kant dice que uno podría llamar la «conciencia de esta ley fundamental un *factum* de la razón» (KpV: 31, 24). La así llamada teoría del *factum* que explica nuestra percepción del carácter obligatorio de la ley moral es, entre otras cosas, una teoría de la justificación. La idea básica es que no puede haber una deducción del imperativo categórico en ningún sentido normal (deductivo),[9] pero la realidad objetiva de la ley moral está «no obstante, firmemente establecida por *sí misma*» (KpV: 47, el énfasis es mío). En nuestra conciencia del

[9] «Por lo tanto, la realidad objetiva de la ley moral no puede ser probada por ninguna deducción» (KpV: 47, 15).

imperativo categórico, la ley moral es dada inmediatamente en su validez incondicional y obligatoria; en este sentido, la teoría del *factum* es una teoría de la auto-evidencia moral. Esta conciencia del imperativo categórico, sin embargo, está determinada por el sentimiento de respeto, esto es, la validez incondicional del imperativo categórico está dada en el sentimiento de respeto. Por lo tanto, es *por medio del sentimiento de respeto que conocemos la validez o carácter obligatorio de la ley moral*.[10] De esta manera, Kant de ningún modo es un racionalista puro, tal como lo representaban Husserl y otros. Más bien, Kant es un intuicionista ético. Un intuicionista ético es alguien que sostiene la postura de que conocemos la validez de la ley moral (el *tú debes* moral) no por medio de algún tipo de razonamiento deductivo (inductivo o abductivo) sino por medio de cierto tipo de auto-evidencia, por un sentimiento. Es importante entender esta tesis de manera correcta: de acuerdo con Kant, *no* es el contenido del imperativo categórico lo que es entendido por el sentimiento de respeto; sabemos *qué* debemos hacer u omitir por medio de la razón y algún tipo de universalización. Por consiguiente, Kant no es un sentimentalista moral. Además, la ley moral no *depende* en sí misma del sentimiento moral de respeto para su validez; no es que la ley moral es válida *porque* tenemos ese sentimiento. De todos modos, lo que sí conocemos mediante el sentimiento de respeto es que debemos actuar moralmente, que la ley moral es categóricamente obligatoria.

1.4. Un argumento del conocimiento kantiano

De lo anterior se sigue, sin embargo, que una computadora no puede tener razón práctica. Para ver esto, tenemos que echar un vistazo al así llamado *argumento del conocimiento* o al argumento a partir del vacío explicativo, expuestos en una u otra versión por Frank Jackson (1982), Joseph Levine (1983) o Thomas Nagel (1974).[11] Se trata de una historia algo intrincada y aquí solo podemos bosquejar la idea principal. Para nuestro propósito, recordemos la historia de María: pensemos en ella como una científica que sabe todo lo que hay que saber sobre colores y su percepción, excepto que ella, que está encerrada en una habitación con solo libros, TV, etc., en blanco y negro, nunca ha visto ningún objeto que no sea blanco o negro. Un día, sin embargo, deja la habitación y ve realmente algo que, según dice, es rojo. Ahora bien, de acuerdo con el fisicalismo, todo lo que existe son objetos naturales (físicos) que son totalmente descriptos y explicados por medio de la física (y posiblemente por la química, la biología y la neurociencia). Si esto fuera cierto, entonces María no conocería una cualidad nueva —y solo sería una cualidad no-representacional de su percepción o de un dato de la sensació — que todavía no ha conocido porque todo lo que hay que conocer sobre colores desde una perspectiva objetiva, en tercera persona, ella lo sabe. Pero hay algo que ella no sabía antes de dejar la habitación, a saber, *cómo es* o *cómo se siente* ver algo rojo, experimentar un cierto *quale* (como dice la jerga). Por lo tanto, hay algo en el mundo que no es físico, i.e. que no es totalmente descriptible por la física. Este algo es la conciencia teniendo experiencias fenoménicas. Entonces, podríamos saber todo lo que hay que saber sobre los hechos físicos o funcionales que conciernen a un estado mental (como por ejemplo tener la experiencia de que algo sea rojo) y todavía no sabríamos todo sobre el estado mental. Por consiguiente, este estado mental no puede ser idéntico a, o reducido a, aquellos hechos físicos o funcionales. Pero *podemos* saber todo acerca de cómo está hecha una computadora y cómo funciona. No existe un *cómo es ser una computadora* y, por lo tanto, a diferencia de los seres para quienes hay una determinada vida interior fenoménica, estar en un estado computacional no significa estar en un estado mental.

Si bien existe cierta disputa sobre qué estados mentales son, o están acompañados por, *qualia*, es claro que los sentimientos son *qualia*. Pero entonces el argumento es evidente: a menos que creamos que las computadoras experimentan *qualia*, no pueden tener razón práctica. En efecto, la razón práctica viene acompañada de constricción práctica a través del sentimiento de respeto. El

[10] Colin, Varner y Zinser (2000: 260) notan que las emociones no solo tienen una función motivacional, sin embargo, no reconocen la función cognitiva con relación a la validez de la ley moral.
[11] Hay algunas diferencias en estos autores, pero asumo que el punto central es el mismo (excepto tal vez en el caso de Levine).

ISSN-e: 2445-0669
DOI 10.7203/REK.3.2.13208

imperativo categórico no puede ser entendido sin este sentimiento. Puesto que las computadoras no tienen sentimientos, y *a fortiori*, ningún sentimiento de respeto, no pueden entender el imperativo categórico.

Este es el argumento básico. Desde un punto de vista kantiano, tres puntos más son importantes. *Primero*, y solo de pasada, debo hacer notar que en su obra tardía (*La metafísica de las costumbres*) Kant desarrolló de manera ulterior su teoría de los sentimientos morales distinguiendo cuatro tipos de predisposiciones morales y, consecuentemente, cuatro sentimientos morales: el sentimiento moral propiamente dicho, la conciencia, el amor a los seres humanos como *amor complacentiae* y el respeto por uno mismo (cf. Schönecker 2010). En relación con cada uno de esos sentimientos, Kant resalta que no hay obligación de tenerlos, pues, en primer lugar, tener esos sentimientos es ya una presuposición necesaria para que tenga sentido el concepto mismo de deber. *Segundo*, para Kant, la razón práctica es la voluntad nouménica que puede tanto conocer como querer el bien, es autónoma, y por lo tanto, una causalidad nouménica. Como he indicado anteriormente, esta es una historia complicada pero los sentimientos morales no pueden ser naturalizados en tanto son producidos por la razón, que tampoco puede ser naturalizada. Por ello, incluso si las computadoras tuvieran sentimientos, no podrían tener el sentimiento de respeto, porque este sentimiento tiene su fuente en la razón, la cual no es una entidad natural (física). *Tercero*, Kant también entiende los sentimientos como *qualia*. Por supuesto, este término no fue utilizado por Kant. Pero él entendía claramente el hecho de que los sentimientos tienen un costado fenoménico que no puede ser comprendido por medio de un conocimiento físico sino que debe ser experimentado. El costado fenoménico de los sentimiento es enfatizado por Kant en su teoría de la belleza (cf. Berger 2019). Los sentimientos como tales, dice Kant en la así llamada *Primera introducción a la Crítica de la facultad de juzgar*, «no pueden de ningún modo ser explicados», más bien «deben ser *sentidos*, no entendidos» [«Man sieht hier leicht, daß Lust oder Unlust, weil sie keine Erkenntnisarten sind, für sich selbst gar nicht können erklärt werden, und gefühlt, nicht eingesehen werden wollen»] (EEKU: 232). En la misma línea, Kant escribe en la *Metafísica de las costumbres* que «el placer y el displacer no pueden por sí mismos ser explicados» (MS: 212).

En cualquier caso, la pureza en la razón pura práctica de ningún modo sugiere que no haya sentimientos involucrados. La pureza de la razón práctica consiste en estar libre de consideraciones sobre la felicidad o el amor propio. En los seres humanos, es la razón pura la que se vuelve práctica sobre la fuerza de los sentimientos morales. Entonces, aun cuando una computadora "tome una decisión" (*por así decirlo*) sobre la base de un algoritmo moral, no tiene idea de lo que es "hacer" (*por así decirlo*): no comprende en absoluto qué es la ley moral como un imperativo categórico. En la terminología de Kant, una computadora puede realizar acciones (*por así decirlo*) conforme al deber. Pero ciertamente no puede realizar acciones por deber. Y ciertamente no tiene conciencia o amor propio; a mi modo de ver, una tesis tal ni siquiera es comprensible.

Antes de pasar a dos problemas de esta tesis kantiana, veamos brevemente un posible argumento ulterior, el *argumento a partir de la facultad de juzgar*. Dada la latitud de muchos deberes éticos, este es obviamente un aspecto importante de la razón práctica. El argumento funciona así: siguiendo a Kant, la facultad de juzgar es la «facultad de pensar el particular como contenido bajo el universal» (KU: 179). Si ya existe una regla, entonces la facultad del juicio es la "facultad de subsumir" algo particular bajo esta regla. Kant llama a esto juicio determinante [*bestimmende Urteilskraft*]. Si todavía hay una regla que ha de ser encontrada para algo particular que no puede ser subsumido bajo una regla adquirida, entonces Kant lo llama *juicio reflexionante* [*reflektierende Urteilskraft*]. Al menos para el juicio determinante, Kant sostiene que no puede haber una regla ulterior. En efecto, si uno «pretendiera mostrar de manera universal cómo se debe subsumir bajo estas reglas, es decir, [cómo se debe] discernir si algo está bajo ellas o no, esto no podría ocurrir de otro modo sino, otra vez, mediante una regla. Pero esta, precisamente por ser una regla, requiere, de

nuevo, una indicación de la facultad de juzgar; y así se pone de manifiesto que si bien el entendimiento es capaz de instrucción y de equipamiento por medio de reglas, la facultad de juzgar es un talento especial que no puede ser enseñado, son solamente ejercido» (KrV: A133/B172).[12] Dicho de otro modo: puede haber meta-reglas acerca de cómo y cuándo aplicar reglas, pero so pena de un círculo vicioso o de un número infinito de reglas, debe haber un punto en el cual la facultad de juzgar actúa sin aplicar una regla. Las computadoras, sin embargo, no tienen más que reglas para trabajar, esto es, nada más que algoritmos (y datos, por supuesto, en relación con los cuales son aplicados). Si la facultad de juzgar es la facultad que no sigue reglas, entonces esta facultad no puede ser algo que una computadora pueda tener. La «carencia de la facultad de juzgar», dice Kant, «es lo que propiamente se llama tontería» (KrV: A133/B172); las computadoras, por lo tanto, son tontas. Pero es difícil ver si este argumento realmente puede ser aceptado. La necesidad de algo como la facultad de juzgar se debe al hecho de que no hay una comprobación completa o definición de todos los conceptos posibles y casos *a priori* de antemano.[13] Pero si una decisión basada en la facultad de juzgar no está basada en una regla, ¿en qué está basada? Mejor que no esté basada en el azar, porque eso es algo que una computadora puede hacer (seguir la regla de elegir al azar). Uno podría pensar que la facultad de juzgar tiene que ver con algo como las intuiciones, pero las intuiciones, como por ejemplo G. E. Moore las entiende, son distintas y no tienen nada que ver con la facultad de juzgar. Las intuiciones —entendidas de manera amplia como estados mentales (como se entienden en psicología moral) de algún modo inconscientes, fuertes y que rápidamente se ven (juzgan) como verdaderos)— muy bien podrían no ser azarosas sino basadas en alguna ponderación (inconsciente, fuerte, rápida) de los bienes, y una ponderación tal podría seguir reglas. Por consiguiente, no estaría preparado para defender el argumento a partir de la facultad de juzgar.

1.5. Yo pienso vs. yo siento

Obviamente, los sujetos morales *quieren* algo y *actúan* sobre la base de sus voliciones. Pero también *piensan*, e incluso si estuviera en discusión cómo el pensamiento, y cuánto de él, está involucrado en sus decisiones morales como tales, no puede estar en discusión que los sujetos morales deban pensar (al menos, en lo que atañe a la cognición del mundo que los rodea en el que quieren y actúan). El conocimiento moral involucra conocimiento no moral sobre el mundo interno y externo. Ahora bien, Kant sostiene célebremente (en la segunda edición de la *Crítica de la razón pura*) que no hay pensamiento y, por lo tanto, conocimiento, del mundo interno y externo sin la autoconciencia. El *yo pienso*, dice, «debe *poder* acompañar todas mis representaciones» (KrV: B131). Qué significa exactamente esto ha sido objeto de dolorosas y largas discusiones entre los especialistas en Kant (cf. Klemme 1996; Rosefeldt 2000).[14] La idea básica parece ser que todo pensamiento involucra la *síntesis* de representaciones como las representaciones de alguien en un juicio, tal que esas representaciones, como el acto de sintetizarlas, pertenecen, y *deben* pertenecer,[15] a un *yo* autoconsciente que siempre puede decir *yo pienso* (esas representaciones). En cualquier caso, si Kant tiene razón, y si es cierto que una computadora no tiene un *yo* que piensa, una computadora no piensa, y no puede pensar, y no es inteligente en el modo en que lo son los seres humanos. A lo sumo, entonces, una computadora (IA) puede imitar la inteligencia, y solo en el mejor caso, imitar el pensamiento moral.

No estoy completamente convencido de este argumento. Me parece cierto que los seres humanos (desarrollados) siempre deben poder pensar *yo pienso* cuando piensan (no que siempre

[12] Las citas en español de la *Crítica de la razón pura* son tomadas de la traducción de Mario Caimi (Buenos Aires, Colihue, 2007) [Nota de la traductora].
[13] «uno se sirve de ciertas notas solo mientras son suficientes para efectuar distinciones, en cambio, nuevas observaciones suprimen algunas [notas], ponen otras en su lugar; así, pues, el concepto no está nunca encerrado en límites seguros» (cf. KrV: B756).
[14] Un ejemplo de cuán difícil puede ser entender bien la posición de Kant es el libro de Kim sobre el "yo pienso" como una proposición empírica (2017). Para una vista rápida, cf. Kitcher (2005).
[15] Friebe (2005: 53) parece pensar que esta lectura es demasiado fuerte.

ISSN-e: 2445-0669
DOI 10.7203/REK.3.2.13208

piensan *yo pienso*, por supuesto). Y si esto es cierto, y es cierto que las computadoras no tienen *yo* (lo cual pienso que de hecho es cierto), entonces las computadoras no piensan en el modo en el que nosotros lo hacemos. Aún podría estar justificado decir que piensan: si el pensamiento teórico (no práctico) es esencialmente el acto de sintetizar contenido que requiere un centro o unidad por medio de la cual este acto es realizado, entonces es posible entender este centro o unidad como la *unidad de control* de una computadora. El punto es que este acto de combinar contenido (síntesis) puede no requerir *Meinigkeit* ni conciencia de sí mismo (*apperception*, como lo llama Kant), como una forma superior de *Meinigkeit*, sino solo una unidad de control. Esta unidad no tiene que ser necesariamente consciente de sí misma porque el contenido de (algo como) las percepciones y de (algo como) los pensamientos o proposiciones no requiere de un *yo* que lo piense e incluso tampoco una forma de *Meinigkeit*. Kant parece haber sostenido que «todo lo que piensa está constituido de la manera como la sentencia de la conciencia de mí mismo [el *yo pienso*] lo declara con respecto a mí» (KrV: B404); «no puedo», afirma Kant, «tener la más mínima representación por una experiencia externa, sino solamente por la conciencia de mí mismo» (KrV: B405). En otras palabras: Kant sostiene que el único modo de concebir la posibilidad del pensamiento en algo distinto a los seres humanos está basado en nuestro propio entendimiento de nosotros mismos. Si esto es cierto, no puedo concebir la posibilidad del pensamiento en una computadora sin pensarlo como un pensamiento de un *yo*. ¿Pero por qué debería ser esto cierto? Esta idea podría ser simplemente una consecuencia del hecho de que todavía no había computadoras.[16]

Entonces puede que quizás exista el pensamiento sin un *yo* consciente de sí mismo que piensa y sin *Meinigkeit*. Cuando se trata de sentimientos, sin embargo, *necesariamente* entramos a un mundo diferente, a saber, al mundo interno. No puede haber sentimientos sin *alguien* que sienta. Dejando de lado la cuestión de si tenemos que distinguir entre sentimientos en sentido estricto y emociones (como intencionales), decir que existe un estado que podríamos describir, de modo preliminar, como "hay un sentimiento", requiere que exista una instancia *a quien* corresponda estar en ese estado (de sentir algo). Una vez más, tal vez exista la posibilidad de concebir contenido cognitivo que no requiera alguien *para quien*, o *de quien*, sea el contenido, pero esto es incomprensible para el caso de los sentimientos.

2. Dos problemas del argumento a partir de los sentimientos morales

Considero que aunque el argumento a partir de los sentimientos morales es fuerte, de todos modos, tiene dos problemas. Primero, ¿qué sucede con las voluntades divinas? Como ya hemos visto, es un elemento importante de la ética de Kant distinguir entre seres divinos y no divinos. Para estos últimos, la ley moral es un imperativo categórico y, por lo tanto, deber y obligación. Para los primeros, no hay obstáculos que la moralidad tenga que superar, ellos tienen una voluntad perfectamente buena. Entonces ¿los seres divinos son en algún sentido máquinas morales? ¿Y no sería cierto que una computadora que siempre siguiera algoritmos morales tendría una voluntad buena perfecta? Bueno, no. Ciertamente, es correcto que un ser divino no puede actuar por deber.[17] Pero, a diferencia de las computadoras, los seres divinos tienen una voluntad. Las computadoras no tienen voluntad, *a fortiori*, no quieren nada por la ley moral. La tesis de que las computadoras no tienen voluntad está respaldada por la tesis de que las voliciones son intencionales,[18] por lo tanto las computadoras no tienen

[16] Ciertamente resta la pregunta sobre el modo en el que las computadoras tienen representaciones. Friebe (2005: 61) tiene razón cuando señala que las representaciones [*Vorstellungen*] como tales son siempre las representaciones de alguien. También podría ser cierto que la propiedad de *ser una representación [Meinigkeit] mía* no implica necesariamente un *yo* en sentido estricto. Sin embargo, permanece la pregunta acerca de si el pensamiento debe ser entendido como una operación que involucra representaciones como algo mental.

[17] Tampoco podría tener las cuatro predisposiciones morales.

[18] Por supuesto que todo esto está sumamente en discusión. Recordemos el largo debate actual sobre el argumento la *habitación china* de Searle. Para un resumen breve y una crítica cf. Gabriel (2018: 95ss).

voluntad. De todos modos, la pregunta acerca de qué significa para una voluntad divina querer y actuar por la ley moral, sin sentimientos morales intermediarios, no es fácil de responder.

El segundo problema con el argumento a partir de los sentimientos morales es simplemente este: *¿las computadoras realmente no pueden tener sentimientos?* Esta, también, es una historia larga y complicada y solo puedo bosquejar el problema y una posible solución. El argumento de que las computadoras quizás puedan tener, o que incluso finalmente tienen, sentimientos dice así: sabemos que *nosotros* tenemos conciencia y sentimientos. Al final del día, no tenemos una historia clara, menos aún comprensible y convincente, que explique cómo esto sucede, cómo realmente puede ser que tengamos una vida interior tal. Sin embargo, si algún tipo de teoría evolutiva naturalista es correcta, sí sabemos que nuestra habilidad de tener eventos mentales se desarrolló a partir de materia inconsciente. Pero si es posible que la mente y sus estados mentales evolucionen de algún modo a partir de la materia, i.e. a partir del cerebro y su encarnación, —y *es* posible toda vez que asumamos que los estados mentales son reales, a pesar de que puedan o no ser reducidos a estados cerebrales— entonces muy bien *podría* ser posible que la mente y sus estados mentales puedan evolucionar de algún modo a partir de una computadora, entendida también como otro ensamble complejo de materia. Está bien, diría yo. Entonces sí, eso es posible si es posible que la mente evolucione a partir de la materia. Pero por todo lo que sabemos es también muy improbable: solo una única célula biológica es *extremadamente* compleja, y más aún el cerebro. Por comparación, una computadora es un objeto muy primitivo, no hay más razones para pensar que una computadora tiene una mente que pensar que la tiene una máquina de coser.[19]

3. Conclusión: submarinos que nadan, planetas que vuelan

Para concluir, quisiera volver al *yo pienso* de Kant. El científico en computación Edsger Dijkstra afirmó célebremente que «la cuestión acerca de si las *máquinas pueden pensar*… es tan relevante como la cuestión acerca de si los *submarinos pueden nadar*» (Russell; Norvig 2016: 1021), o bien, uno podría agregar, si los aviones pueden volar. Supongo que el punto de Dijkstra es que los submarinos *pueden* nadar, *por supuesto*, i.e. moverse a través del agua a pesar del hecho de que no nadan como los peces y que los aviones pueden volar a pesar del hecho de que se mueven a través del aire sin mover las alas hacia arriba y hacia abajo, o como sea. Siguiendo la analogía de Dijkstra, parece sensato sostener que las computadoras piensan a pesar del hecho de que no piensan del modo en el que nosotros lo hacemos; no hay un *yo* y, sin embargo, piensan. Pero esta analogía entre computadoras que piensan, submarinos que nadan o aviones que vuelan es engañosa. Como siempre, todo depende de cómo se definan las palabras, como "nadar", "volar" o "pensar". Si uno define "nadar" como "moverse a través del agua usando sus miembros, aletas o colas", entonces los submarinos no nadan, pero ¿por qué lo traduciríamos así? Definir una palabra —o explicar qué es para una cosa ser lo que es— presupone de alguna manera reconocer qué es esencial para esa cosa; para hacer eso, sin embargo, uno necesita casos paradigmáticos (cf. Damschen; Schönecker 2019). Ahora bien, uno podría estar de acuerdo —quizás en la línea del funcionalismo— en que las computadoras, sobre la base de determinados *inputs*, realizan determinadas operaciones que causan determinados *outputs* y, tomando en consideración solo el *output* —calcular, jugar al ajedrez, manejar un auto, componer música— uno está tentado a pensar que las computadoras piensan (y también que los seres humanos piensan en el modo en el que lo hacen las computadoras). La pregunta acerca de cómo y por qué medios se nada no es crucial para el concepto de nadar. Pero la diferencia entre un ser que piensa *yo pienso*, o al menos entre uno que siente *yo siento*, o experimenta *yo quiero*, y una

[19] Uno podría también, dicho sea de paso, dar vueltas al asunto y sostener que la existencia de estados mentales prueba que algo está mal con las teorías evolutivas darwinistas. Thomas Nagel, entre otros, ha hecho eso recientemente (2012). Para una visión crítica del realismo moral de Nagel, cf. Schmidt (2018).

ISSN-e: 2445-0669
DOI 10.7203/REK.3.2.13208

máquina que no tiene tal conciencia de sí es tan enorme que esos términos (*pensar*, *sentir*, *querer*) no deberían usarse para seres que no tienen un yo. Decir que una computadora piensa o siente es como decir que una planeta *vuela* solo porque se mueve a través del espacio.

Bibliography

EEKU = *Erste Einleitung in die Kritik der Urteilskraft*

GMS = *Grundlegung zur Metaphysik der Sitten*

KpV = *Kritik der praktischen Vernunft*

KrV = *Kritik der reinen Vernunft*

MS = *Metaphysik der Sitten*

BERGER, L.: *Kants Philosophie des Schönen. Eine kommentarische Interpretation zu den §§ 1-22 der „Kritik der Urteilskraft* (en prensa), 2019.

BONNEFON, J.-F.; Shariff, A.; Rahwan, I.: "Autonomous Vehicles Need Experimental Ethics: Are we ready for Utilitaran Cars?", *arXiv:1510.03346 [cs.CY]* (12/10/2015).

COLIN, A.; VARNER, G.; ZINSER, J.: "Prolegomena to any future artificial moral agent", *Journal of Experimental & Theoretical Artificial Intelligence* 12, 3 (2000) 251-261.

DAMSCHEN, G.; SCHÖNECKER, D.: *Selbst philosophieren. Ein Methodenbuch*, 3ª ed., Berlín/Boston, de Gruyter, 2019.

FRIEBE, C.: *Theorie des Unbewußten: Eine Deutung der Metapsychologie Freuds aus transzendentalphilosophischer Perspektive*, Würzburg, Königshausen & Neumann, 2005.

GABRIEL, M.: *Der Sinn des Denkens*, Berlín, Ullstein Verlag, 2018.

HUSSERL, E.: "Kritik der Kantischen Ethik," en: *Vorlesungen über Ethik und Wertlehre 1908-1914* (Husserliana XXXVIII), Dordrecht/Boston/London, 1988, 402-418.

JACKSON, F.: "Epiphenomenal Qualia", *Philosophical Quarterly* 32 (1982) 127-136.

KIM, H.: *Zur Empirizität des „Ich denke in Kants Kritik der reinen Vernunft*, Würzburg, Königshausen & Neumann, 2017.

KITCHER, P.: "Ich denke", en WILLASCHEK, M.; STOLZENBERG, J.; MOHR, G.; BACIN, S. (eds.), *Kant-Lexikon*, 2, Berlín/Boston, de Gruyter, 2015, 1074-1079.

KLEMME, H. F.: *Kants Philosophie des Subjekts. Systematische und entwicklungsgeschichtliche Untersuchungen zum Verhältnis von Selbstbewßtsein und Selbsterkenntnis*, Hamburg, Meiner Verlag, 1996.

LEVINE, J: "Materialism and Qualia: The Explanatory Gap", *Pacific Philosophical Quarterly* 64 (1983) 354-361.

MORELAND, J. P.; WILLIAM L. C.: *Philosophical Foundations for a Christian Worldview*, Illinois, InterVarsity Press, 2003.

NAGEL, T.: "What it is like to be a bat", *Philosophical Review* 83 (1974) 435-450.

_____: *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False*, Oxford, Oxford University Press, 2012.

POTHAST, U.: *Die Unzulänglichkeit der Freiheitsbeweise. Zu einigen Lehrstücken aus der neueren Geschichte von Philosophie und Recht*, Frankfurt, Suhrkamp, 1987.

ROSEFELDT, T.: *Das logische Ich. Kant über den Gehalt des Begriffes von sich selbst*, Berlín, philo Verlag, 2000.

RUSSEL, S. J.; NORVIG, P.: *Artificial Intelligence. A Modern Approach*, Londres, Pearson Education Limited, 2016.

SCHADOW, S.: *Achtung für das Gesetz: Moral und Motivation bei Kant*, Berlín/Boston, de Gruyter, 2012.

SCHMIDT, E. E.: "*The Dilemma of Moral Naturalism in Nagel's* Mind and Cosmos", *Ethical Perspectives* 25 (2018) 203-231.

_____: *Kants Begriff der Demütigung in der "Kritik der praktischen Vernunft"*, Master-Arbeit, Universität Siegen, 2013.

SCHMIDT, E. E.; SCHÖNECKER D.: "Kant's Ground-Thesis. On Dignity and Value in the Groundwork", *The Journal of Value Inquiry* 52 (2018) 81-95.

SCHÖNECKER, D.: *Kants Begriff transzendentaler und praktischer Freiheit. Eine entwicklungs-geschichtliche Studie* (en colaboración con Stefanie Buchenau y Desmond Hogan), Berlín, Walter de Gruyter (Kant-Studien Ergänzungshefte), 2005.

_____: "Kant über Menschenliebe als moralische Gemütsanlage", *Archiv für Geschichte der Philosophie* (en colaboración con Alexander Cotter, Magdalena Eckes, Sebastian Maly) 2 (2010) 133-175.

_____: "Das gefühlte Faktum der Vernunft. Skizze einer Interpretation und Verteidigung", *Deutsche Zeitschrift für Philosophie* 1 (2013a) 91-107.

_____: "Kant's Moral Intuitionism. The Fact of Reason and Moral Predispositions", *Kant Studies Online* (2013b) 1-38.

SCHÖNECKER, D.; WOOD, A.: *Kant's Groundwork for the Metaphysics of Morals. A Commentary*, Cambridge, Harvard University Press, 2015.

ULGEN, O.: "Kantian Ethics in the Age of Artificial Intelligence and Robotics", *Questions of International Law*, Zoom-in 43 (2017) 59-83.

WAKWFIELD, J.: "South Korean university boycotted over 'killer robots'", BBC News *(online)*, 2018, https://www.bbc.com/news/technology-43653648.

WALLACH, W.; ALLEN, C.: *Moral machines: Teaching robots right from wrong*, Nueva York, Oxford University Press, 2009.