Por qué es casi seguro que tu mente no está (ni estará nunca) en un ordenador*

Jesús Zamora Bonilla
jpzb@fsof.uned.es

En este artículo ofrezco algunos argumentos para desmontar dos ideas recurrentes en la reciente literatura transhumanista: la de que podremos alcanzar una especie de «vida eterna virtual», grabando el contenido de nuestra mente en un ordenador, y la tesis de que es muy probable que, de hecho, estemos viviendo ya en el interior de una especie de «simulación informática». Un principio importante de cualquier epistemología social razonable dice que el porcentaje de ideas que *son* absurdas, entre las ideas que *suenan* absurdas, es extremadamente elevado. Naturalmente, un montón de ideas de las que sonaban absurdas han terminado mostrándose acertadas (por ejemplo, la idea de la evolución de especies diferentes a partir de un antepasado común, la idea de que la tierra es un planeta que gira en torno a una estrella, la idea de que la materia está formada por *átomos*, etc.), pero por cada una de estas «victorias del ingenio contra el sentido común», miles de afirmaciones absurdas han existido y existirán. Esto significa que no te estás comportando como un estúpido reaccionario cuando tildas instintivamente una idea como «estúpida» si ves claramente que contradice al sentido común, sino solo que tu cerebro está poniendo en práctica un sano escepticismo. «Las afirmaciones extravagantes requieren pruebas extraordinarias», y tu escepticismo natural solo tiene permitido batirse en retirada cuando se empiezan a presentar dichas pruebas.

Desgraciadamente, bastantes filósofos, en su noble y legítima tarea de poner a prueba los límites del sentido común, han trastabillado a menudo con la recomendación que acabo de poner en negrita, interpretándola más o menos como que «las afirmaciones extraordinarias requieren pruebas extravagantes». Más de un milenio al servicio de la teología, engendrando una serie inacabable de ar-

* Este artículo se enmarca en el proyecto de investigación FFI2017-89639-P.

gumentos extraordinariamente sagaces y sutiles sobre la existencia y las propiedades de dios, de los *ángeles,* de los demonios, de los santos y de las almas, ha dejado seguramente en algunos de nuestros filósofos más capaces una tendencia imborrable a tomarse un poquitín demasiado en serio algunas conjeturas extravagantes. Tampoco tenemos que olvidar que, en lo que se refiere a cuestiones de hecho, las «pruebas extraordinarias» no pueden provenir más que de hallazgos empíricos, y en especial de la confirmación de predicciones acertadas pero muy implausibles. Un argumento meramente verbal, por muy sofisticado que parezca, no puede nunca servir de algo que no sea una tautología. Por tanto, la probabilidad de que un filósofo apoye una idea-que-suena-absurda solo porque resulta *sexy*, más que porque haya razones válidas para apoyarla, tiende a ser probablemente mayor que lo que imaginas.

¿PODRÁ GRABARSE NUESTRA MENTE EN UN ORDENADOR?

La idea de que la mente es al cerebro como el *software* es al *hardware* de un ordenador (una idea compartida tanto por el computacionalismo como por el conexionismo) ha llevado a muchos a pensar en la posibilidad de que ese *software* podría ser traducido a una serie de ceros y unos, igual que podemos hacer con los sonidos de un concierto, y archivarlos en un dispositivo informático adecuado. La idea es especialmente popular entre los defensores del llamado *transhumanismo* (véase, p. e., Seung, 2012). Seguramente el principal atractivo de esta idea es que, como el mito del espíritu, nos hace una promesa de inmortalidad, una promesa que en este caso puede revestirse de verosimilitud a la vista del progreso tecnológico. Por desgracia, como argumentaré en este artículo, es una promesa exactamente igual de vacía.

El precursor de la tesis de que «la mente es información» es nada menos que Aristóteles. Según el viejo filósofo griego, cuando la mente percibe una cosa, lo que hace es «recibir la *forma*» *(morphé)* de la cosa, *pero sin recibir su materia*, de modo semejante a como el sello deja su marca en la cera caliente. El pensamiento *(noûs)* sería también similar en esto a la percepción sensorial, solo que, en ese caso, en vez de percibir la «forma sensible» del objeto (color, textura, sonido...), lo que se *graba* en nuestra mente su *forma inteligible*, o sea, su esencia, lo que la cosa es (y de este modo, nosotros lo *entendemos* como una mesa, o un caballo, etc.). La mente, por tanto, sería para Aristóteles algo así como lo que puede recibir la *forma* de todas las cosas, sin necesidad de que esas cosas se introduzcan «materialmente» en ella, al contrario que lo que sucede en otras funciones corporales, como la digestión o la respiración, en las que sí que debemos incorporar a nosotros la *materia* correspondiente.

Sería un anacronismo identificar sin más la noción contemporánea de *información* con el concepto aristotélico de *forma*, pero la relación etimológica entre

ambos es mucho más que meramente casual. Una concepción más puesta al día de la noción aristotélica de *forma* podríamos identificarla con la idea de *estructura*: no es que el sonido de la campana sea una «cualidad sensible» que en nuestra mente se pueda separar del bronce del que está hecha, sino que es un *patrón matemático* de ondas sonoras, que hacen vibrar los *órganos* de nuestro oído según ese mismo patrón, y que allí se convierte en otro patrón estructuralmente relacionado con el primero, pero esta vez no es un patrón de vibraciones sonoras, sino de estímulos eléctricos transmitidos por las neuronas. Nuestra mente consistiría en algo así como en la *suma de todos esos patrones matemáticos* de corrientes eléctricas que circulan por nuestro cerebro (más el patrón de las conexiones establecidas entre las neuronas, el «conectoma» del que habla Seung). Y un patrón matemático no es, en el fondo, más que lo que llamamos *información*.

El problema, naturalmente, es pasar de esta premisa aparentemente trivial («nuestros estados mentales consisten en algún tipo de información contenida en el cerebro») a las siguientes conclusiones:

1. será técnicamente posible *copiar y archivar* esa información en un dispositivo artificial, y
2. la información así grabada servirá de alguna manera para «reproducir» nuestra mente.

Voy a exponer brevemente los principales argumentos por los que creo que no es razonable aceptar ninguna de estas conclusiones (argumentos en gran medida inspirados en Aaronson, 2013).

En primer lugar, la cantidad de datos que deberían extraerse es tan descomunal (un cerebro adulto contiene cientos de *billones* de sinapsis, p. e.), que, aunque el improbable advenimiento de los *computadores cuánticos* resolviera el problema de dónde almacenar esa información, es difícil imaginar que alguna vez se desarrolle una «tecnología que permita observar sin error» todas y cada una de esas conexiones en un cerebro humano vivo (e incluso en uno muerto).

En segundo lugar, aunque tuviéramos un escáner lo bastante potente como para registrar todas y cada una de nuestras sinapsis, está el problema, mencionado en el apartado anterior, de que «ignoramos casi por completo el código» (o, mejor dicho, la suma de miles o millones de códigos) que convierte esos datos en recuerdos, pensamientos o decisiones específicas. Es como si tenemos un CD con música grabada, pero no tenemos ningún programa que pueda transformarlo en sonidos; o como si inventásemos una máquina para grabar conversaciones que estaban teniendo lugar hace 30.000 años, pero ignorásemos el lenguaje en el que esas personas hablaban: en ese caso, no entenderemos una palabra de lo que estaban diciendo. Este problema es muy grave, porque nuestra mente no consiste solo, ni siquiera principalmente, en tener *almacenados* unos ciertos recuerdos, rasgos de carácter, etc., sino sobre todo en estar recordando, estar decidiendo, estar escuchando, etc., es decir, en estar llevando a cabo actividades mentales.

El problema del código consiste, pues, en que, incluso si poseyéramos toda la información sobre cómo está el cerebro en un *instante* determinado, tendríamos que hacer correr esa información con algún programa (o programas) para que se convirtiese en un *proceso* que pudiera tomarse por una actividad mental, y esos programas, de momento, no parece que tengamos ni la menor idea de cómo descubrirlos.

En tercer lugar, incluso aunque hubiésemos logrado superar las dos dificultades previas (cómo escanear el cerebro con el detalle suficiente, y qué programas aplicar a los datos así recogidos para que nuestra «mente informática» logre ejecutar una auténtica actividad mental), queda el problema de que la actividad mental que lleva a cabo un cerebro vivo depende de una continua interacción con el resto de su organismo y con su entorno. Una mente archivada en un ordenador y «funcionando» dirigida por un programa que replicase las funciones cerebrales, ¿estaría sintiendo calor o frío?, *¿en* qué podría pensar?, *¿qué* decisiones podría tomar? A falta de contacto con una realidad física de la que recibir *nueva* información y sobre la cual actuar, quizá lo máximo a lo que podría llegar una mente así sería a estar en algo parecido a un sueño. Como defienden numerosos especialistas, las capacidades y estados mentales de cualquier animal, incluidos los seres humanos, no dependen solo de lo que sucede en el cerebro, sino también de las peculiaridades de nuestros organismos y de los seres que nos rodean, en interacción con los cuales han evolucionado durante millones de años. La mente es «corpórea», «extendida». O como decía el filósofo José Ortega y Gasset: «yo soy yo y mi circunstancia». Una buena introducción a esta teoría es Rowlands (2010).

Quizá esta *última* dificultad pudiera superarse proporcionando a esa mente que está grabada en un ordenador una especie de «circunstancia orteguiana artificial». Pero, para que la experiencia de esa persona fuera verdaderamente realista, eso exigiría replicar no solo su propia mente, sino todo el universo, o al menos algo tan complejo como el universo, en el que los organismos pudieran funcionar hasta el *último* detalle como lo hacen en el mundo real, o de algún modo equivalente. Además, tampoco está claro qué tipo de «vida eterna» sería esa: *¿una* sucesión de episodios desconectados entre sí, cada vez que el programa se pone en marcha?, *¿una única* experiencia que se repite una y otra vez? *¿O* un futuro completamente abierto (aunque esto *último* no soy capaz de ver cómo podría ser programado en un ordenador)?

Otra posible *solución* sería que, en vez de proporcionar a la mente informática un organismo y un entorno simulados, quizá podría volver a grabarse el contenido de la mente en un cerebro *virgen*, por así decir. Pero si la posibilidad de *escanear* un cerebro vivo con todo el detalle necesario parece remotísima, la de *fabricar* un cerebro que fuese una réplica neurona por neurona, sinapsis por sinapsis, de otro cerebro que hubiéramos escaneado previamente creo que está fuera de los límites de cualquier tecnología posible.

Se han señalado también otros problemas del proyecto de «hacer una copia informática de la mente». P. e., *¿la* copia sería realmente yo, o sería *solo* una copia? *¿Serían* todas las copias idénticas a mí en caso de que se hiciesen varias? *¿Sería* aceptable destruir el cuerpo que ha sido *escaneado* una vez que se ha fabricado una copia perfecta de *él*? Estos problemas, de todas formas, presuponen que es factible hacer ese *copiado*, lo que, por los argumentos que he dado más arriba, me parece algo totalmente inaccesible tecnológicamente, así que no los discutiré aquí.

POR QUÉ ES CASI SEGURO QUE *NO* VIVIMOS EN UNA SIMULACIÓN

Desde luego, ni Nick Bostrom, ni Elon Musk, han sido los primeros en dar popularidad a la tesis de que el mundo que experimentamos puede ser una especie de ficción. En la tradición de la filosofía occidental, tanto Platón como Descartes son famosos por sugerir algo así, el primero con su «mito de la caverna», y el segundo con su «genio maligno», pero la idea tiene aún una tradición más antigua en Oriente (p. e., el «velo de Maya»). La popularidad actual de la conjetura de que vivimos en una realidad ilusoria debe mucho, por supuesto, a la creciente industria de los juegos de ordenador y a los aparatos de realidad virtual, así como a su difusión en películas como *Matrix* o *Desafío total*. Podríamos decir que, hacia el principio de este siglo, el mundo estaba maduro para recibir algún intento de dignificación intelectual de esta moda. *¿Y* qué podría ser mejor que una prueba lógica o matemática? Por supuesto, si tenemos en cuenta que gran parte de la audiencia potencial de este argumento son frikis de la tecnología, ese tipo de prueba será mucho más aceptable que un balbuceo cuasi ininteligible sobre la ontología de los simulacros elaborado por un pedante filósofo continental. Nick Bostrom, por entonces un joven y prometedor filósofo analítico con una fuerte base lógica y matemática, tuvo *éxito* en proporcionar justo lo que el mundo estaba esperando.

El argumento de Bostrom, muy resumido, es el siguiente. O bien es extraordinariamente improbable que la humanidad (u otra forma de vida inteligente) evolucione hasta alcanzar la capacidad de crear «perfectas simulaciones cósmicas» (quizá porque tiendan a extinguirse antes de ello), o bien existe algo (p. e., un tabú cultural) que impedirá llevar a cabo esas simulaciones, o bien las dos hipótesis anteriores son falsas y, por tanto, en algún momento futuro, alguna civilización lo suficientemente sofisticada decidirá implementar un número astronómico de tales simulaciones. Parece que las dos primeras hipótesis pueden ser descartadas como muy implausibles, consideradas como leyes sin ninguna excepción, y por lo tanto es prácticamente seguro que, en algún momento de la historia del universo, alguna civilización alcanzará la capacidad de realizar «perfectas simulaciones cósmicas» casi sin límite (pensemos, p. e., en ordenadores cuánticos, cuyos bits capaces de llevar a cabo trillones de operaciones simultá-

neamente), tal vez con el objetivo de *observar* y *experimentar* lo que sucede en dichas simulaciones, o tal vez por pura diversión. Ahora bien, esto implica que, si existen o existirán billones de universos perfectamente simulados, la probabilidad de que el universo que estamos observando sea «el real» es ridículamente pequeña en comparación con la probabilidad de que sea uno de esos billones de simulaciones.

Antes de entrar a analizar los pasos de esta argumentación, reflexionemos sobre un argumento que tiene cierta semejanza. Como dijo una vez Bertrand Russell, es estrictamente imposible refutar la conjetura de que el mundo ha empezado a existir hace justo cinco minutos en el estado en el que se encontraba en ese preciso momento. *¿Implica* esto que es igual de probable que el universo observable haya comenzado a existir hace justo cinco minutos, y que haya comenzado a existir en el Big Bang, más o menos hace 13.500 millones de años? Quizá estemos tentados a responder que no, pero imaginemos que, en vez de considerar solo esas dos opciones, producimos una serie astronómicamente grande de conjeturas alternativas: que el mundo haya empezado a existir hace cinco minutos, o hace cinco minutos y un nanosegundo, o hace cinco minutos y dos nanosegundos, etc. Hay un número astronómicamente alto de posibles momentos en los que el mundo podría haber empezado a existir «tal como era entonces», y, por lo tanto, parece que la conjetura de que empezó a existir *justo* en el Big Bang, y no después, tiene una probabilidad microscópicamente baja de ser verdadera.

Nuestra inteligencia se retuerce (con buenas razones) contra esta conclusión, porque la enorme magnitud del número de conjeturas estúpidas que hemos producido artificialmente no hace que cada una de ellas sea ni un microgramo menos absurda de lo que era cuando solo teníamos *dos* conjeturas (una de ellas absurda, y la otra, no). Y la combinación de un número astronómico de conjeturas absurdas parece que no deja de ser bastante absurda. Pensamos, simplemente, que es extremadamente más probable que el universo observable empezara a existir con el Big Bang, que no que empezase a hacerlo en cualquier momento posterior «tal como era justo entonces». Y nuestra principal razón para pensar así es que «las leyes de la física no tendrían mucho sentido en caso contrario».

Espero que este *último* argumento haya servido para quitarle un poco del encanto a la «magia de los grandes números» en la que la tesis de Bostrom quiere fundamentarse. Seguramente, la principal razón por la que estamos ante un argumento falaz tiene que ver con su propio contenido, más que con su estructura formal. El razonamiento de Bostrom se basa en una simple extrapolación a partir del progreso histórico de la tecnología; una extrapolación que es bastante naíf: al contrario que Bostrom, mi impresión es que lo que es bastante probable es que el progreso tecnológico no pueda conducir a ciertas situaciones imaginarias simplemente proyectando las tendencias históricas de los *últimos* siglos o décadas, sino que más bien lo probable es que cada civilización termine su desarrollo tecnológico en una especie de «meseta» a partir de la cual no sea posible un progre-

so significativamente mayor (aunque, por supuesto, en muchos casos podemos estar todavía muy lejos de las «mesetas» más elevadas). Después de todo, si la tecnología pudiera progresar indefinidamente, el universo debería estar lleno de señales de civilizaciones mucho más avanzadas que la nuestra.

Mi principal contraargumento, de todas formas, es que los datos sobre los que se basa la extrapolación de Bostrom relativa al progreso tecnológico son datos sobre *nuestro propio mundo*. Por tanto, si nuestro mundo no fuese *real*, sino una *simulación*, entonces *no habría absolutamente ninguna razón para pensar que los datos obtenidos de un mundo falso son relevantes y representativos de lo que ocurriría en un mundo real*. Dicho de otro modo: *si el argumento de Bostrom es correcto, eso implica que no podemos usar una de las premisas en las que está basado*. Por ejemplo, quizá nuestro mundo es una simulación, pero una llevada a cabo en un universo *real* cuyas leyes físicas (quizá muy diferentes de las que pensamos que el nuestro obedece) solo permiten a los habitantes *reales* de ese universo la posibilidad de crear un número muy pequeño de simulaciones, no un número astronómicamente alto (recuérdese que el argumento de Bostrom no solo necesita que en el futuro puedan llevarse a cabo algunas simulaciones, sino un número enorme de ellas).

¿Qué es lo que falla, entonces, en el trilema de Bostrom? Sospecho que, entre sus tres opciones iniciales, la de que las «simulaciones cósmicas *perfectas*» son físicamente imposibles es la opción más probable de todas: si una simulación así fuese posible, entonces la gente que habita esa simulación tendría que ser capaz de construir su propio mundo ficticio *perfecto* (pues si no pudiera, entonces ya habría algo en lo que se diferencian de los seres *reales*), pero entonces la gente de este tercer mundo podrían crear sus propias simulaciones cósmicas perfectas, cuyos habitantes podrían crear otras, etc. Pero parece que hay límites informacionales que impiden que un número de «simulaciones perfectas *anidadadas*» puedan existir, ya que el mundo *real* originario, en el que la primera simulación (y, por ende, las demás) están siendo implementadas, solo puede dedicar una cantidad *finita* de recursos a ella, y, por lo tanto, la cantidad de información que la simulación contendrá será necesariamente limitada, y en tal caso, no podrá haber una serie infinita de simulaciones anidadas (lo que, como digo, sería necesario para que cualquiera de esas simulaciones fuese *perfecta*, esto es, indistinguible de un mundo *real*). Las simulaciones cósmicas, por lo tanto, no pueden por principio ser *perfectas* (o sea, dar toda la impresión de que contienen un universo físico *completo*, hasta sus *últimos* detalles), salvo, de nuevo, que las leyes del universo *real* sean tan diferentes a las del nuestro que sencillamente no podamos inferir absolutamente nada sobre todo ello.

Una razón adicional por la que es muy dudoso que seamos el producto de una simulación informática es la siguiente: las simulaciones se llevan a cabo normalmente con el objetivo de *observar* algunos de sus resultados (de hecho, esta es la idea que tiene Bostrom en mente cuando imagina a las futuras civilizaciones construyendo esos mundos simulados). Esto significa que las entidades

de las que la simulación está compuesta, o los sucesos que tienen lugar dentro de ella, deben tener alguna conexión física con algo que sea utilizado como un interfaz entre la simulación y los sujetos que la observan desde fuera. Pero las líneas causales de nuestro propio universo parecen estar *cerradas*, en el sentido de que ninguna energía ni información puede escapar de nuestro universo para ser transferidas a esa imaginaria «interfaz». Por ejemplo, acontecimientos que hayan ocurrido en un pasado muy lejano y que, a causa de la segunda ley de la termodinámica, no han dejado trazas que permitan conocerlos *ahora* (p. e., *¿era* macho o hembra el *último* dinosaurio que murió?), esos acontecimientos no solo es imposible averiguar *ahora* si han sucedido o no, sino que tampoco podrían averiguarlos los creadores de la simulación «cuando acaben los tiempos», es decir, cuando dejen de *hacer correr* el programa. Y si los estaban observando «en tiempo real» (si, p. e., están observándote ahora) entonces para hacerlo deberían haber ejercido *alguna interacción física detectable*, una especie de *milagro* (desde nuestro punto de vista interior en el «mundo simulado»), información causal que se *escapa* del mundo, lo que parece que no se observa de ningún modo. Esta *última* discusión me lleva a sugerir, por cierto, la *única* forma en la que creo que los defensores de la hipótesis de la simulación podrían realmente intentar verificarla: no mediante argumentos lógicos o filosóficos (que siempre son altamente dudosos cuando se refieren a escenarios especulativos), sino mediante la confirmación *empírica* de que existe alguna interfaz física entre los constructores de la simulación y nosotros. A falta de esa confirmación, su tesis no deja de ser una más de los millones de tesis absurdas que ha inventado la humanidad a lo largo de la historia.

Terminaré ofreciendo una *última* razón para sospechar sobre la validez del argumento de la simulación, una que es probablemente más profunda desde el punto de vista filosófico: la cuestión es que el argumento depende de una concepción muy naíf del conocimiento, al entenderlo meramente como una especie de «representación mental del mundo externo»; esta concepción lleva a plantear la cuestión como si la analogía más relevante consistiera en cómo distinguir un cuadro original de Velázquez, digamos, de una buena copia (la *simulación*). Pero el caso es que el conocimiento no es, en su forma más básica, una *representación*, pese a que a veces, o a menudo, utilicemos representaciones con el fin de obtener conocimientos. El conocimiento es más bien un tipo de *actividad práctica, material* (como respirar, caminar o reproducirse) que llevamos a cabo *interactuando* con las cosas que nos rodean. La realidad material no es primariamente algo *externo* que podemos tratar de *conocer* y distinguir de las *ilusiones*, sino que es un elemento intrínseco de la actividad en la que consiste conocer. Como vimos que decían Rowlands y otros, el conocimiento es esencialmente «conocimiento incorporado» *(embodied)*, y está formado más por *prácticas* inferenciales que por *representaciones* mentales. Incluso la realidad virtual es, en *último* término, nada más que una *porción* de nuestro universo material, con la que también tenemos que aprender a interactuar de una manera determinada.

REFERENCIAS

S. AARONSON: *Quantum computing since Democritus*. Cambridge, CUP, 2013.

N. BOSTROM: «Are you living in a computer simulation?», *Philosophical Quarterly*, vol. 53, n.º 211 (2003), pp. 243-255.

M. ROWLANDS: *The new science of the mind: from extended mind to embodied phenomenology*. Cambridge (Ma., EE.UU.), MIT Press, 2010.

S. SEUNG: *Connectome: how the brain's wiring makes us who we are*. Houghton, Boston, 2012.

JESÚS ZAMORA BONILLA (Madrid, 1963) es catedrático de Filosofía de la Ciencia en la Facultad de Filosofía de la UNED, y coordinador del máster en Periodismo y Comunicación Científica de esa universidad. Es doctor en Filosofía y en Ciencias Económicas, y autor de numerosos libros y artículos académicos, así como de varias obras literarias. Entre sus últimos libros de filosofía están *Cuestión de protocolo. Ensayos de metodología de la ciencia* (Madrid: Tecnos, 2005) y *Sacando consecuencias. Una filosofía para el siglo XXI* (Madrid: Tecnos, 2017).