¿VIVE EN UNA SIMULACIÓN DE COMPUTADORA?

POR NICK BOSTROM
Facultad de Filosofía, Universidad de Oxford
Publicado en *Philosophical Quarterly* (2003) Vol. 53, núm. 211, págs. 243-255.
[www.simulation-argument.com]
versión pdf: [PDF]

ABSTRACTO

Este artículo sostiene que *al menos una* de las siguientes proposiciones es cierta: (1) es muy probable que la especie humana se extinga antes de alcanzar una etapa "posthumana"; (2) es extremadamente improbable que cualquier civilización posthumana ejecute un número significativo de simulaciones de su historia evolutiva (o variaciones de la misma); (3) es casi seguro que vivamos en una simulación por computadora. De ello se desprende que la creencia de que existe una posibilidad significativa de que algún día seamos posthumanos que realicen simulaciones de ancestros es falsa, a menos que estemos viviendo actualmente en una simulación. También se analizan otras consecuencias de este resultado.

I. INTRODUCCIÓN

Muchas obras de ciencia ficción, así como algunos pronósticos de tecnólogos y futurólogos serios, predicen que en el futuro estarán disponibles enormes cantidades de potencia informática. Supongamos por un momento que estas predicciones son correctas. Una cosa que las generaciones posteriores podrían hacer con sus computadoras superpoderosas es ejecutar simulaciones detalladas de sus antepasados o de personas como sus antepasados. Debido a que sus computadoras serían tan poderosas, podrían ejecutar muchas de esas simulaciones. Supongamos que estas personas simuladas son conscientes (como lo serían si las simulaciones fueran lo suficientemente detalladas y si cierta posición ampliamente aceptada en la filosofía de la mente fuera correcta). Entonces podría darse el caso de que la gran mayoría de mentes como la nuestra no pertenezcan a la raza original sino a personas simuladas por los descendientes avanzados de una raza original. Entonces es posible argumentar que, si este fuera el caso, sería racional pensar que probablemente estemos entre las mentes simuladas en lugar de entre las biológicas

originales. Por lo tanto, si no pensamos que estamos viviendo actualmente en una simulación por computadora, no tenemos derecho a creer que tendremos descendientes que ejecutarán muchas de esas simulaciones de sus antepasados. Esa es la idea básica. El resto de este documento lo explicará con más detalle. Por lo tanto, si no pensamos que estamos viviendo actualmente en una simulación por computadora, no tenemos derecho a creer que tendremos descendientes que ejecutarán muchas de esas simulaciones de sus antepasados. Esa es la idea básica. El resto de este documento lo explicará con más detalle. Por lo tanto, si no pensamos que estamos viviendo actualmente en una simulación por computadora, no tenemos derecho a creer que tendremos descendientes que ejecutarán muchas de esas simulaciones de sus antepasados. Esa es la idea básica. El resto de este documento lo explicará con más detalle.

Aparte del interés que esta tesis puede tener para quienes se dedican a la especulación futurista, también existen recompensas más puramente teóricas. El argumento proporciona un estímulo para formular algunas preguntas metodológicas y metafísicas , y sugiere analogías naturalistas con ciertas concepciones religiosas tradicionales, que algunos pueden encontrar divertidas o estimulantes.

La estructura del documento es la siguiente. Primero, formulamos una suposición que necesitamos importar de la filosofía de la mente para comenzar la discusión. En segundo lugar, consideramos algunas razones empíricas para pensar que ejecutar una gran cantidad de simulaciones de mentes humanas estaría dentro de la capacidad de una civilización futura que ha desarrollado muchas de esas tecnologías que ya se puede demostrar que son compatibles con leyes físicas conocidas y limitaciones de ingeniería. Esta parte no es filosóficamente necesaria pero proporciona un incentivo para prestar atención al resto. Luego sigue el núcleo del argumento, que hace uso de alguna teoría de probabilidad simple, y una sección que brinda apoyo para un principio de indiferencia débil que emplea el argumento.Por último, discutimos algunas interpretaciones de la disyunción, mencionada en abstracto, que forma la conclusión del argumento de la simulación.

II. LA ASUNCIÓN DE INDEPENDENCIA DEL SUSTRATO

Un supuesto común en la filosofía de la mente es el de *la independencia* del *sustrato* . La idea es que los estados mentales pueden sobrevenir en cualquiera de una amplia clase de sustratos físicos. Siempre que un sistema implemente el tipo correcto de estructuras y procesos computacionales, puede asociarse con experiencias conscientes. No es una propiedad esencial de la conciencia que se implemente en redes neuronales biológicas basadas en carbono

dentro de un cráneo: los procesadores basados en silicio dentro de una computadora podrían, en principio, hacer el truco también.

Se han dado argumentos a favor de esta tesis en la literatura, y aunque no es del todo incontrovertible, aquí la daremos por hecho.

El argumento que presentaremos, sin embargo, no depende de ninguna versión muy fuerte de funcionalismo o computacionalismo. Por ejemplo, no necesitamos suponer que la tesis de la independencia del sustrato es *necesariamente*cierto (ya sea analítica o metafísicamente), solo que, de hecho, una computadora que ejecute un programa adecuado sería consciente. Además, no necesitamos suponer que para crear una mente en una computadora sería suficiente programarla de tal manera que se comporte como un humano en todas las situaciones, incluida la aprobación de la prueba de Turing, etc. Solo necesitamos la suposición más débil que sería suficiente para la generación de experiencias subjetivas que los procesos computacionales de un cerebro humano se replican estructuralmente con un detalle adecuadamente detallado, como en el nivel de las sinapsis individuales. Esta versión atenuada de la independencia del sustrato es bastante aceptada.

Los neurotransmisores, los factores de crecimiento nervioso y otras sustancias químicas que son más pequeñas que una sinapsis claramente juegan un papel en la cognición y el aprendizaje humanos. La tesis de la independencia del sustrato no es que los efectos de estos productos químicos sean pequeños o irrelevantes, sino que afectan la experiencia subjetiva solo a *través de* su influencia directa o indirecta en las actividades computacionales. Por ejemplo, si no puede haber diferencia en la experiencia subjetiva sin que también haya una diferencia en las descargas sinápticas, entonces el detalle requerido de la simulación está en el nivel sináptico (o superior).

III. LOS LÍMITES TECNOLÓGICOS DE LA COMPUTACIÓN

En nuestra etapa actual de desarrollo tecnológico, no tenemos ni el hardware ni el software lo suficientemente potentes para crear mentes conscientes en las computadoras. Pero se han dado argumentos persuasivos en el sentido de que *si* el progreso tecnológico continúa sin cesar , estas deficiencias eventualmente se superarán. Algunos autores argumentan que esta etapa puede estar a solo unas décadas de distancia. [1] Sin embargo, los propósitos actuales no requieren suposiciones sobre la escala de tiempo. El argumento de la simulación funciona igualmente bien para aquellos que piensan que se necesitarán cientos de miles de años para alcanzar una etapa de civilización "posthumana", donde la humanidad ha adquirido la mayoría de las capacidades tecnológicas que actualmente se pueden

demostrar que son consistentes con las leyes físicas y con limitaciones de material y energía.

Una etapa tan madura de desarrollo tecnológico permitirá convertir planetas y otros recursos astronómicos en computadoras enormemente poderosas. Actualmente, es difícil confiar en algún límite superior de la potencia informática que pueda estar disponible para las civilizaciones posthumanas. Como todavía nos falta una "teoría del todo", no podemos descartar la posibilidad de que fenómenos físicos novedosos, no permitidos en las teorías físicas actuales, puedan ser utilizados para trascender esas restricciones [2] que en nuestra comprensión actual imponen límites teóricos a el procesamiento de la información que se puede obtener en un determinado trozo de materia. Podemos con mucha mayor confianza establecer *menores*límites en la computación posthumana, asumiendo sólo mecanismos que ya se entienden. Por ejemplo, Eric Drexler ha esbozado un diseño para un sistema del tamaño de un terrón de azúcar (excluyendo la refrigeración y la fuente de alimentación) que realizaría 10^{21} instrucciones por segundo. [3] Otro autor da una estimación aproximada de 10^{42} operaciones por segundo para una computadora con una masa del orden de un planeta grande. [4] (Si pudiéramos crear computadoras cuánticas, o aprender a construir computadoras a partir de materia nuclear o plasma, podríamos acercarnos a los límites teóricos. Seth Lloyd calcula un límite superior para una computadora de 1 kg de $5 * 10^{50}$ operaciones lógicas por segundo realizado en ~ 10^{31} bits. [5] Sin embargo, es suficiente para nuestros propósitos usar la estimación más conservadora que presupone solo principios de diseño conocidos actualmente).

La cantidad de potencia informática necesaria para emular una mente humana también se puede estimar de forma aproximada. Una estimación, basada en lo costoso desde el punto de vista computacional que es replicar la funcionalidad de un trozo de tejido nervioso que ya hemos entendido y cuya funcionalidad se ha replicado *in silico* , la mejora del contraste en la retina, arroja una cifra de ~ 10^{14} operaciones por segundo para todo el cerebro humano. [6] Una estimación alternativa, basada en el número de sinapsis en el cerebro y su frecuencia de disparo, da una cifra de ~ 10^{16} -10^{17} operaciones por segundo. [7] Es posible que se requieran aún más si queremos simular en detalle el funcionamiento interno de las sinapsis y los árboles dendríticos. Sin embargo, es probable que el sistema nervioso central humano tenga un alto grado de redundancia en la microescala para compensar la falta de fiabilidad y el ruido de sus componentes neuronales. Por lo tanto, cabría esperar una ganancia de eficiencia sustancial cuando se utilizan procesadores no biológicos más fiables y versátiles.

La memoria no parece ser una restricción más estricta que la capacidad de procesamiento. [8] Además, dado que el ancho de banda sensorial humano máximo es ~ ^{108} bits por segundo, la simulación de todos los eventos sensoriales incurre en

un costo insignificante en comparación con la simulación de la actividad cortical. Por lo tanto, podemos utilizar la potencia de procesamiento necesaria para simular el sistema nervioso central como una estimación del costo computacional total de simular una mente humana.

Si el entorno se incluye en la simulación, esto requerirá potencia de cálculo adicional; cuánto depende del alcance y la granularidad de la simulación. Simular todo el universo hasta el nivel cuántico es obviamente inviable, a menos que se descubra una física radicalmente nueva. Pero para obtener una simulación realista de la experiencia humana, se necesita mucho menos, solo lo que se requiere para garantizar que los humanos simulados, que interactúan de manera humana normal con su entorno simulado, no noten ninguna irregularidad. La estructura microscópica del interior de la Tierra se puede omitir con seguridad. Los objetos astronómicos distantes pueden tener representaciones muy comprimidas: la verosimilitud debe extenderse a la estrecha banda de propiedades que podemos observar desde nuestro planeta o la nave espacial del sistema solar. En la superficie de la Tierra*ad hoc* . Lo que ve a través de un microscopio electrónico debe parecer poco sospechoso, pero generalmente no tiene forma de confirmar su coherencia con partes no observadas del mundo microscópico. Las excepciones surgen cuando diseñamos deliberadamente sistemas para aprovechar fenómenos microscópicos no observados que operan de acuerdo con principios conocidos para obtener resultados que podemos verificar de forma independiente. El caso paradigmático de esto es una computadora. Por lo tanto, la simulación puede necesitar incluir una representación continua de computadoras hasta el nivel de elementos lógicos individuales. Esto no presenta ningún problema, ya que nuestra potencia informática actual es insignificante para los estándares posthumanos.

Además, un simulador posthumano tendría suficiente potencia de cálculo para realizar un seguimiento de los estados de creencias detallados en todos los cerebros humanos en todo momento. Por lo tanto, cuando vio que un humano estaba a punto de hacer una observación del mundo microscópico, pudo completar la simulación con suficiente detalle en el dominio apropiado según sea necesario. Si ocurriera algún error, el director podría editar fácilmente los estados de cualquier cerebro que se haya dado cuenta de una anomalía antes de que arruine la simulación. Alternativamente, el director podría retroceder unos segundos y volver a ejecutar la simulación de una manera que evite el problema.

Por lo tanto, parece plausible que el principal costo computacional en la creación de simulaciones que son indistinguibles de la realidad física para las mentes humanas en la simulación reside en la simulación de cerebros orgánicos hasta el nivel neuronal o subneuronal. [9] Si bien no es posible obtener una estimación muy exacta del costo de una simulación realista de la historia humana, podemos usar ~ 10^{33} - 10^{36} operaciones como una estimación aproximada [10]. A medida que

ganemos más experiencia con la realidad virtual, comprenderemos mejor los requisitos computacionales para hacer que esos mundos parezcan realistas para sus visitantes. Pero en cualquier caso, incluso si nuestra estimación se desvía en varios órdenes de magnitud, esto no importa mucho para nuestro argumento. Notamos que una aproximación aproximada de la potencia computacional de una computadora de masa planetaria es 10^{42} operaciones por segundo, y eso supone solo diseños nanotecnológicos ya conocidos, que probablemente están lejos de ser óptimos. Una sola computadora de este tipo podría simular toda la historia mental de la humanidad (llamen a esto una *simulación de ancestros*) utilizando menos de una millonésima parte de su potencia de procesamiento durante un segundo. Una civilización posthumana puede eventualmente construir una cantidad astronómica de tales computadoras. Podemos concluir que el poder de cómputo disponible para una civilización posthumana es suficiente para ejecutar una gran cantidad de simulaciones de ancestros, incluso si asigna solo una pequeña fracción de sus recursos a ese propósito. Podemos sacar esta conclusión incluso dejando un margen de error sustancial en todas nuestras estimaciones.

- Λασ civilizaciones posthumanas tendrían suficiente poder de cómputo para ejecutar una gran cantidad de simulaciones de ancestros, incluso cuando solo usan una pequeña fracción de sus recursos para ese propósito.

IV. EL NÚCLEO DEL ARGUMENTO DE LA SIMULACIÓN

La idea básica de este artículo se puede expresar a grandes rasgos de la siguiente manera: si existiera una posibilidad sustancial de que nuestra civilización llegara alguna vez a la etapa posthumana y ejecutara muchas simulaciones de ancestros, entonces ¿cómo es que no estás viviendo en tal simulación?

Desarrollaremos esta idea en un argumento riguroso. Introduzcamos la siguiente notación:

f_P: Fracción de todas las civilizaciones tecnológicas de nivel humano que sobreviven para alcanzar una etapa posthumana.

$\overline{N}$: Número medio de simulaciones de antepasados realizadas por una civilización posthumana.

$\overline{H}$: Número promedio de individuos que han vivido en una civilización antes de que alcance una etapa posthumana.

La fracción real de todos los observadores con experiencias de tipo humano que viven en simulaciones es entonces

$$f_{sim} = \frac{f_P \overline{N} \overline{H}}{(f_P \overline{N} \overline{H}) + \overline{H}}$$

Escribir para la fracción de civilizaciones posthumanas que están interesadas en ejecutar simulaciones de antepasados (o que contienen al menos algunos individuos que están interesados en eso y tienen recursos suficientes para ejecutar un número significativo de tales simulaciones), y para el número promedio de antepasados simulaciones llevadas a cabo por civilizaciones tan interesadas, tenemos $f_I \quad \overline{N_I}$

$$\overline{N} = f_I \overline{N_I}$$

y así:

$$f_{sim} = \frac{f_P f_I \overline{N_I}}{(f_P f_I \overline{N_I}) + 1} \qquad (*)$$

Debido al inmenso poder de cómputo de las civilizaciones posthumanas, es extremadamente grande, como vimos en la sección anterior. Al inspeccionar (*) podemos ver que *al menos una* de las siguientes tres proposiciones debe ser verdadera: $\overline{N_I}$

(1) $f_P \approx 0$
(2) $f_I \approx 0$
(3) $f_{sim} \approx 1$

V. UN PRINCIPIO DE INDIFERENCIA BLANDA

Podemos dar un paso más y concluir que, condicionado a la verdad de (3), la credibilidad de uno en la hipótesis de que uno está en una simulación debe estar cerca de la unidad. De manera más general, si supiéramos que una fracción *x* de todos los observadores con experiencias de tipo humano viven en simulaciones, y no tenemos ninguna información que indique que nuestras propias experiencias particulares son más o menos probables que otras experiencias de tipo humano. se han implementado *in vivo en* lugar de *en machina* , entonces nuestra credibilidad de que estamos en una simulación debería ser igual a *x* :

$$Cr(SIM \mid f_{sim} = x) = x \qquad (\#)$$

Este paso está sancionado por un principio de indiferencia muy débil. Distingamos dos casos. El primer caso, que es el más fácil, es donde todas las mentes en cuestión son como la suya en el sentido de que son exactamente cualitativamente idénticas a la suya: tienen exactamente la misma información y las mismas experiencias que usted tiene. El segundo caso es donde las mentes son "similares" entre sí sólo en el sentido vago de ser el tipo de mentes que son típicas de las criaturas humanas, pero son cualitativamente distintas entre sí y cada una tiene un conjunto distinto de experiencias. Sostengo que incluso en el último caso, donde las mentes son cualitativamente diferentes, el argumento de la simulación todavía funciona, siempre que no tenga información que se refiera a la cuestión de cuáles de las diversas mentes se simulan y cuáles se implementan biológicamente.

En la literatura se ha dado una defensa detallada de un principio más fuerte, que implica la postura anterior para ambos casos como casos especiales triviales. [11] El espacio no permite una recapitulación de esa defensa aquí, pero podemos resaltar una de las intuiciones subyacentes llamando nuestra atención sobre una situación análoga de un tipo más familiar. Supongamos que el x % de la población tiene una determinada secuencia genética S dentro de la parte de su ADN comúnmente designada como "ADN basura". Supongamos, además, que no hay manifestaciones de S (salvo las que aparecerían en un ensayo de genes) y que no hay correlaciones conocidas entre tener Sy cualquier característica observable. Entonces, con toda claridad, a menos que usted ha tenido su ADN secuenciado, es racional para asignar un crédito de x % a la hipótesis de que tiene S . Y esto es así bastante independientemente del hecho de que las personas que tienen S tienen cualitativamente diferentes mentes y experiencias de las personas que no tienen S . (Son diferentes simplemente porque todos los humanos tienen experiencias diferentes entre sí, no por ningún vínculo conocido entre S y el tipo de experiencias que uno tiene).

El mismo razonamiento es válido si S no es la propiedad de tener una determinada secuencia genética, sino la propiedad de estar en una simulación, asumiendo solo que no tenemos información que nos permita predecir cualquier diferencia entre las experiencias de las mentes simuladas y las de las mentes simuladas. mentes biológicas originales.

Cabe destacar que el suave principio de indiferencia expresado por (#) prescribe indiferencia solo entre hipótesis sobre qué observador eres, cuando no tienes información sobre cuál de estos observadores eres. En general, no prescribe indiferencia entre hipótesis cuando se carece de información específica sobre cuál de las hipótesis es verdadera. En contraste con Laplacean y otros principios de indiferencia más ambiciosos, es por lo tanto inmune a la paradoja de Bertrand y

predicamentos similares que tienden a plagar los principios de indiferencia de alcance ilimitado.

A los lectores familiarizados con el argumento del día del juicio final [12] les puede preocupar que el suave principio de indiferencia invocado aquí sea el mismo supuesto que es responsable de hacer despegar el argumento del día del juicio final, y que el carácter contrario a la intuición de algunas de las implicaciones de este último incrimina o arroja duda sobre la validez del primero. No es así. El argumento del día del juicio final se basa en una premisa *mucho* más fuerte y controvertida, a saber, que uno debe razonar como si fuera una muestra aleatoria del conjunto de todas las personas que alguna vez habrá vivido (pasado, presente y futuro) *aunque sepamos que están viviendo a principios del siglo XXI*en lugar de en algún momento del pasado distante o del futuro. El principio de indiferencia suave, por el contrario, se aplica solo a los casos en los que no tenemos información sobre a qué grupo de personas pertenecemos.

Si las probabilidades de apostar brindan alguna orientación a la creencia racional, también puede valer la pena considerar que si todos apostaran sobre si están en una simulación o no, entonces si las personas usan el principio suave de la indiferencia y, en consecuencia, colocan su dinero al estar en una simulación si saben que ahí es donde están casi todas las personas, entonces casi todos ganarán sus apuestas. Si apuestan a *no* estar en una simulación, casi todos perderán. Parece mejor que se preste atención al suave principio de indiferencia.

Además, se puede considerar una secuencia de situaciones posibles en las que una fracción cada vez mayor de todas las personas vive en simulaciones: 98%, 99%, 99,9%, 99,9999%, etc. A medida que uno se acerca al caso límite en el que *todos* están en una simulación (del cual uno puede inferir *deductivamente* que uno mismo está en una simulación), es plausible exigir que la credibilidad que uno asigna a estar en una simulación se acerque gradualmente al caso límite de certeza completa de una manera coincidente.

NOSOTROS. INTERPRETACIÓN

La posibilidad representada por la proposición (1) es bastante sencilla. Si (1) es cierto, es casi seguro que la humanidad no alcanzará un nivel posthumano; pues prácticamente ninguna especie en nuestro nivel de desarrollo se vuelve posthumana, y es difícil ver alguna justificación para pensar que nuestra propia especie será especialmente privilegiada o protegida de futuros desastres. Condicional a (1), por lo tanto, debemos dar un alto crédito a *DOOM* , la hipótesis de que la humanidad se extinguirá antes de alcanzar un nivel posthumano:

$$Cr(DOOM \mid f_p \approx 0) \approx 1$$

Uno puede imaginar situaciones hipotéticas eran tenemos pruebas tales como el conocimiento triunfaría de . Por ejemplo, si descubrimos que estamos a punto de ser golpeados por un meteoro gigante, esto podría sugerir que hemos tenido una mala suerte. Entonces podríamos asignar un crédito a *DOOM* más grande que nuestra expectativa de la fracción de civilizaciones a nivel humano que no logran alcanzar la posthumanidad. En el caso real, sin embargo, parece que nos faltan pruebas para pensar que somos especiales en este sentido, para bien o para mal. f_P

La proposición (1) no implica por sí misma que es probable que nos extingamos pronto, solo que es poco probable que alcancemos una etapa posthumana. Esta posibilidad es compatible con que permanezcamos en, o algo por encima, de nuestro nivel actual de desarrollo tecnológico durante mucho tiempo antes de extinguirnos. Otra forma de que (1) sea cierto es si es probable que la civilización tecnológica colapse. Entonces, las sociedades humanas primitivas podrían permanecer en la Tierra indefinidamente.

Hay muchas formas en que la humanidad podría extinguirse antes de llegar a la posthumanidad. Quizás la interpretación más natural de (1) es que es probable que nos extingamos como resultado del desarrollo de alguna tecnología poderosa pero peligrosa. [13] Un candidato es la nanotecnología molecular, que en su etapa de madurez permitiría la construcción de nanobots autorreplicantes capaces de alimentarse de suciedad y materia orgánica, una especie de bacteria mecánica. Estos nanobots, diseñados con fines maliciosos, podrían causar la extinción de toda la vida en nuestro planeta. [14]

La segunda alternativa en la conclusión del argumento de la simulación es que la fracción de civilizaciones posthumanas que están interesadas en ejecutar la simulación de ancestros es insignificante. Para que (2) sea cierto, debe haber una fuerte *convergencia* entre los cursos de las civilizaciones avanzadas. Si el número de simulaciones de ancestros creadas por las civilizaciones interesadas es extremadamente grande, la rareza de tales civilizaciones debe ser correspondientemente extrema. Prácticamente ninguna civilización posthumana decide utilizar sus recursos para ejecutar un gran número de simulaciones de ancestros. Además, prácticamente todas las civilizaciones posthumanas carecen de individuos que tengan suficientes recursos e interés para ejecutar simulaciones de antepasados; o bien han hecho cumplir de manera confiable leyes que impiden que dichas personas actúen según sus deseos.

¿Qué fuerza podría provocar tal convergencia? Se puede especular que todas las civilizaciones avanzadas se desarrollan a lo largo de una trayectoria que lleva al reconocimiento de una prohibición ética de ejecutar simulaciones de antepasados debido al sufrimiento que se inflige a los habitantes de la simulación. Sin embargo, desde nuestro punto de vista actual, no está claro que crear una raza humana sea

inmoral. Por el contrario, tendemos a ver la existencia de nuestra raza como un gran valor ético. Además, la convergencia en una visión ética de la inmoralidad de ejecutar simulaciones de ancestros no es suficiente: debe combinarse con la convergencia en una estructura social de toda la civilización que permita que las actividades consideradas inmorales sean efectivamente prohibidas.

Otro posible punto de convergencia es que casi todos los posthumanos individuales en prácticamente todas las civilizaciones posthumanas se desarrollan en una dirección en la que pierden sus deseos de ejecutar simulaciones de ancestros. Esto requeriría cambios significativos en las motivaciones que impulsan a sus predecesores humanos, porque ciertamente hay muchos humanos a quienes les gustaría ejecutar simulaciones de ancestros si pudieran permitírselo. Pero tal vez muchos de nuestros deseos humanos sean considerados tontos por cualquiera que se convierta en posthumano. Tal vez el valor científico de las simulaciones de ancestros para una civilización posthumana sea insignificante (lo cual no es demasiado inverosímil dada su insondable superioridad intelectual), y tal vez los posthumanos consideren las actividades recreativas simplemente como una forma muy ineficaz de obtener placer, que se puede obtener de manera mucho más barata mediante la estimulación directa de los centros de recompensa del cerebro. Una conclusión que se desprende de (2) es que las sociedades posthumanas serán muy diferentes de las sociedades humanas: no contendrán agentes independientes relativamente ricos que tengan toda la gama de deseos similares a los humanos y sean libres de actuar en consecuencia.

La posibilidad expresada por la alternativa (3) es conceptualmente la más intrigante. Si vivimos en una simulación, entonces el cosmos que estamos observando es solo una pequeña parte de la totalidad de la existencia física. La física en el universo donde se encuentra la computadora que ejecuta la simulación puede o no parecerse a la física del mundo que observamos. Si bien el mundo que vemos es en cierto sentido "real", no está ubicado en el nivel fundamental de la realidad.

Es posible que las civilizaciones simuladas se vuelvan posthumanas. Luego, pueden ejecutar sus propias simulaciones de antepasados en computadoras poderosas que construyen en su universo simulado. Tales computadoras serían "máquinas virtuales", un concepto familiar en informática. (Los web-applets de Java script, por ejemplo, se ejecutan en una máquina virtual, una computadora simulada, dentro de su escritorio). Las máquinas virtuales se pueden apilar: es posible simular una máquina que simula otra máquina, y así sucesivamente, en muchos pasos arbitrarios de iteración. Si continuamos creando nuestras propias simulaciones de ancestros, esto sería una fuerte evidencia en contra de (1) y (2), y por lo tanto tendríamos que concluir que vivimos en una simulación. Además, tendríamos que sospechar que los posthumanos que ejecutan nuestra simulación son ellos mismos seres simulados; y sus creadores, a su vez,

Por tanto, la realidad puede contener muchos niveles. Incluso si es necesario que la jerarquía toque fondo en algún momento (el estado metafísico de esta afirmación es algo oscuro), puede haber espacio para una gran cantidad de niveles de realidad, y el número podría ir aumentando con el tiempo. (Una consideración que cuenta en contra de la hipótesis de niveles múltiples es que el costo computacional de los simuladores del nivel del sótano sería muy alto. Simular incluso una sola civilización posthumana podría ser prohibitivamente costoso. Si es así, deberíamos esperar que nuestra simulación termine cuando estamos a punto de convertirnos en posthumanos).

Aunque todos los elementos de tal sistema pueden ser naturalistas, incluso físicos, es posible establecer algunas analogías vagas con las concepciones religiosas del mundo. De alguna manera, los posthumanos que ejecutan una simulación son como dioses en relación con las personas que habitan la simulación: los posthumanos crearon el mundo que vemos; son de inteligencia superior; son "omnipotentes" en el sentido de que pueden interferir en el funcionamiento de nuestro mundo incluso en formas que violen sus leyes físicas; y son "omniscientes" en el sentido de que pueden controlar todo lo que sucede. Sin embargo, todos los semidioses, excepto aquellos en el nivel fundamental de la realidad, están sujetos a las sanciones de los dioses más poderosos que viven en niveles inferiores.

Una mayor reflexión sobre estos temas podría culminar en una *teogonía naturalista* que estudiaría la estructura de esta jerarquía, y las limitaciones impuestas a sus habitantes por la posibilidad de que sus acciones en su propio nivel puedan afectar el trato que reciben de los habitantes de niveles más profundos. Por ejemplo, si nadie puede estar seguro de que está en el sótano, entonces todo el mundo tendría que considerar la posibilidad de que sus acciones sean recompensadas o castigadas, basándose quizás en criterios morales, por sus simuladores. Una vida futura sería una posibilidad real. Debido a esta incertidumbre fundamental, incluso la civilización del sótano puede tener una razón para comportarse éticamente. El hecho de que tenga tal razón para el comportamiento moral, por supuesto, se sumaría a la razón de todos los demás para comportarse moralmente, y así sucesivamente, en un círculo verdaderamente virtuoso. Uno podría obtener una especie de imperativo ético universal,

Además de las simulaciones de ancestros, también se puede considerar la posibilidad de simulaciones más selectivas que incluyan solo un pequeño grupo de humanos o un solo individuo. El resto de la humanidad serían entonces zombis o "personas de las sombras": humanos simulados solo a un nivel suficiente para que las personas completamente simuladas no noten nada sospechoso. No está claro cuánto más barato sería simular las personas en las sombras que las personas reales. Ni siquiera es obvio que sea posible que una entidad se comporte de manera indistinguible de un humano real y, sin embargo, carezca de experiencia

consciente. Incluso si hay simulaciones tan selectivas, no debe pensar que está en una de ellas a menos que crea que son mucho más numerosas que las simulaciones completas.

También existe la posibilidad de que los simuladores reduzcan ciertas partes de la vida mental de los seres simulados y les proporcionen recuerdos falsos del tipo de experiencias que normalmente habrían tenido durante el intervalo omitido. Si es así, uno puede considerar la siguiente (descabellada) solución al problema del mal: que no hay sufrimiento en el mundo y que todos los recuerdos del sufrimiento son ilusiones. Por supuesto, esta hipótesis se puede considerar seriamente solo en aquellos momentos en los que actualmente no está sufriendo.

Suponiendo que vivamos en una simulación, ¿cuáles son las implicaciones para nosotros los humanos? A pesar de las observaciones anteriores, las implicaciones no son tan radicales. Nuestra mejor guía sobre cómo nuestros creadores posthumanos han elegido configurar nuestro mundo es el estudio empírico estándar del universo que vemos. Las revisiones de la mayor parte de nuestras redes de creencias serían bastante leves y sutiles, en proporción a nuestra falta de confianza en nuestra capacidad para comprender las costumbres de los posthumanos. Por lo tanto, bien entendida, la verdad de (3) no debería tener tendencia a hacernos "volvernos locos" o impedir que nos dediquemos a nuestros asuntos y hagamos planes y predicciones para el mañana. La principal importancia empírica de (3) en el momento actual parece residir en su papel en la conclusión tripartita establecida anteriormente. [15] Podemos esperar que (3) sea cierto ya que eso disminuiría la probabilidad de (1), aunque si las restricciones computacionales hacen probable que los simuladores terminen una simulación antes de que alcance un nivel posthumano, entonces nuestra mejor esperanza sería que (2) es verdad.

Si aprendemos más sobre las motivaciones posthumanas y las limitaciones de recursos, tal vez como resultado de nuestro desarrollo para convertirnos en posthumanos, entonces la hipótesis de que somos simulados llegará a tener un conjunto mucho más rico de implicaciones empíricas.

VII. CONCLUSIÓN

Una civilización "posthumana" tecnológicamente madura tendría una enorme potencia informática. Basado en este hecho empírico, el argumento de la simulación muestra que *al menos una* de las siguientes proposiciones es verdadera: (1) La fracción de civilizaciones a nivel humano que alcanzan una etapa posthumana es muy cercana a cero; (2) La fracción de civilizaciones posthumanas que están interesadas en ejecutar simulaciones de ancestros es muy cercana a cero; (3) La fracción de todas las personas con nuestro tipo de experiencias que viven en una simulación es muy cercana a uno.

Si (1) es cierto, entonces es casi seguro que nos extinguiremos antes de llegar a la posthumanidad. Si (2) es cierto, entonces debe haber una fuerte convergencia entre los cursos de las civilizaciones avanzadas para que prácticamente ninguna contenga individuos relativamente ricos que deseen ejecutar simulaciones de ancestros y sean libres de hacerlo. Si (3) es cierto, es casi seguro que vivamos en una simulación. En el oscuro bosque de nuestra ignorancia actual, parece sensato repartir la credibilidad de uno aproximadamente de manera uniforme entre (1), (2) y (3).

A menos que ahora estemos viviendo en una simulación, es casi seguro que nuestros descendientes nunca ejecutarán una simulación de ancestros.

Agradecimientos

Agradezco a mucha gente por sus comentarios, y especialmente a Amara Angelica, Robert Bradbury, Milan Cirkovic, Robin Hanson, Hal Finney, Robert A. Freitas Jr., John Leslie, Mitch Porter, Keith DeRose , Mike Treder, Mark Walker, Eliezer Yudkowsky y varios árbitros anónimos.

[Página de inicio académica de Nick Bostrom: www.nickbostrom.com]
[Más sobre el argumento de la simulación: www.simulation-argument.com]

---

[1] Véase, por ejemplo, KE Drexler,*Engines of Creation: The Coming Era of Nanotechnology*,Londres, Forth Estate, 1985; N. Bostrom, "¿Cuánto tiempo antes de la superinteligencia?" *Revista Internacional de Estudios de Futuros,* vol. 2, (1998); R. Kurzweil, *La era de las máquinas espirituales: cuando las computadoras superan la inteligencia humana* , Nueva York, Viking Press, 1999; H. Moravec, *Robot: Mera Machine to Transcendent Mind* , Oxford University Press, 1999.

[2] Como ellímite deBremermann-Bekensteiny el límite del agujero negro (HJ Bremermann, "Requisitos mínimos de energía de la transferencia de información y la computación".*International Journal of Theoretical Physics*21: 203-217 (1982); JDBekenstein, "Entropy content y flujo de información en sistemas con energía limitada. "*Physical Review*D 30: 1669-1679 (1984); A. Sandberg," La física de lossuperobjetos deprocesamiento deinformación: La vida diaria entre los cerebros de Júpiter ".*Revista de evolución y tecnología*, vol. 5 (1999)).

[3] KE Drexler,*Nanosystems: Molecular Machinery, Manufacturing, and Computation*,Nueva York, John Wiley & Sons, Inc., 1992.

[4] RJ Bradbury, "MatrioshkaBrains". *Manuscrito de trabajo*(2002), http://www.aeiveos.com/~bradbury/MatrioshkaBrains/MatrioshkaBrains.html.

[5] S. Lloyd, "Límites físicos definitivos para la computación". *Nature*406 (31 de agosto): 1047-1054 (2000).

[6] H. Moravec,*Mind Children*,Harvard Universidad Prensa (1989).

[7] Bostrom (1998), op. cit.

[8] Véanse las referencias en las notas a pie de página anteriores.

[9] A medida que construimos más computadoras y más rápidas, el costo de simular nuestras máquinas podría llegar a dominar el costo de simular sistemas nerviosos.

[10] 100 mil millones de humanos50 años / humano30 millones desegundos/año[10^{14}, 10^{17}] operaciones en cada cerebro humano por segundo[10^{33}, 10^{36}] operaciones.

[11] En p. Ej. N. Bostrom, "El argumento del Juicio Final, Adam & Eve, UN $^{++}$ y Quantum Joe". *Synthese* 127 (3): 359 - 387 (2001); y más plenamente en mi libro *Anthropic Bias: Observation Selection Effects in Science and Philosophy* , Routledge, Nueva York, 2002.

[12] Véase, por ejemplo, J. Leslie, "Is the End of the World Nigh? " *Philosophical Quarterly* 40, 158: 65-72 (1990).

[13] Consulte mi artículo "Riesgos existenciales: análisis de escenarios de extinción humana y peligros relacionados". *Revista de Evolución y Tecnología,*vol. 9 (2001) para un estudio y análisis de las amenazas presentes y futuras previstas para la supervivencia humana.

[14] Véase, por ejemplo, Drexler (1985) op cit., Y RA Freitas Jr., "Some Limits to GlobalEcophagy byBiovorous Nanoreplicators, with Public Policy Recommendations". *Zyvexpreimpresión* de abril (2000), http://www.foresight.org/NanoRev/Ecophagy.html.

[15] Para algunas reflexiones de otro autor sobre las consecuencias de (3), que fueron provocadas por una versión anterior de este artículo que circuló en forma privada, consulte R. Hanson, "Cómo vivir en una simulación". *Revista de Evolución y Tecnología*, vol. 7 (2001).