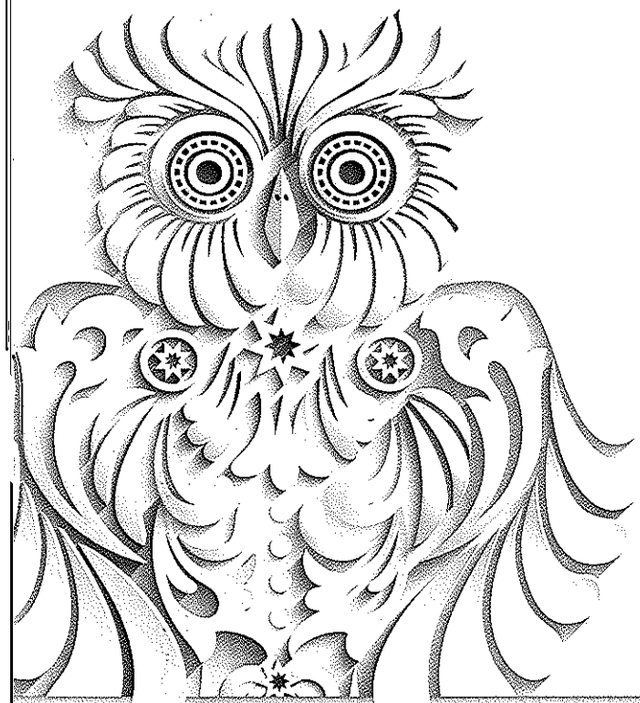
Nick Bostrom

erinteligencia

caminos, peligros, estrategias



1011 0110100 01100111 0110010

imes

ellers

"Recomiendo encarecidamente este libro"

Bill Gates

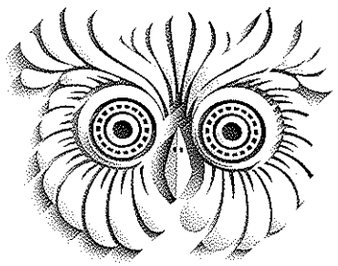
El cerebro humano tiene unas capacidades de las que carecen otros cerebros animales. Es debido a estas capacidades distintivas que nuestra especie ocupa una posición dominante. Otros animales tienen músculos más fuertes o garras más afiladas, pero nosotros tenemos cerebros más inteligentes.

Si los cerebros artificiales llegaran algún día a superar a los cerebros humanos en inteligencia general, entonces esta nueva inteligencia llegaría a ser muy poderosa. De igual manera que el destino de los gorilas depende ahora más de los humanos que de ellos mismos, también el destino de nuestra especie pasaría a depender de las acciones de la superinteligencia artificial.

Pero nosotros tenemos una ventaja: podemos hacer el primer movimiento. ¿Será posible construir una IA seminal o preparar las condiciones iniciales para que podamos sobrevivir a la explosión de inteligencia? ¿Cómo podríamos lograr una detonación controlada?

Este libro profundamente ambicioso y original traza su camino cuidadosamente a través de un terreno vasto y de dificultad prohibitiva. Pero su escritura es tan lúcida que, de alguna manera, lo hace parecer fácil. Tras una travesía absorbente que nos lleva a pensar sobre los límites de la condición humana y el futuro de la vida inteligente, encontramos que el trabajo de Nick Bostrom supone toda una reconceptualización del tema esencial de nuestro tiempo.

01101110 01100011 01100101



Superinteligencia

INTERNACIONAL
LIBROS
MIGUEL CONCHA S.A.
IMPORTACIÓN DIRECTA
LIBROS TÉCNICOS - CIENTÍFICOS - MEDICINA
DICCIONARIOS
Alférez Real 1414, Providencia, Santiago

2016
c.3

SUPERINTELIGENCIA

Caminos, peligros, estrategias

NICK BOSTROM
Director, Future of Humanity Institute
Profesor, Facultad de Filosofía y Oxford Martin School,
Universidad de Oxford

UNIVERSIDAD DIEGO PORTALES
BIBLIOTECA CENTRAL
Reg.: 167810

Todo Está En Los Libros
EDITORIAL

Publicación en inglés por OXFORD University Press
Título Original: SUPERINTELLIGENCE, Paths, Dangers, Strategies
Derechos de autor ©2014 by Nick Bostrom
All rights reserved
ISBN: 978-0-19-967811-2

Superintelligence was originally published in English in 2014. This translation is published by arrangement with Oxford University Press. Teell Editorial is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuracies or ambiguities in such translation or for any losses caused by reliance thereon.

Superinteligencia fue publicada originalmente en inglés en 2014. Esta traducción se publica con acuerdo con Oxford University Press. Teell Editorial solamente es responsable de la traducción del trabajo original y Oxford University Press no tendrá ninguna responsabilidad de cualquier error, omisión, ambigüedad o descuido en dicha traducción o por cualquier pérdida causada por la dependencia de la misma.

EDITADO POR TEELL EDITORIAL, S.L.
Primera edición en español: Teell Editorial, S.L. 2016
www.teelleditorial.com
Traductor e Introducción: Marcos Alonso

Diseño cubierta: www.uypdesign.com

Reservados todos los derechos. Queda rigurosamente prohibida, sin la autorización escrita de los titulares del "Copyright", bajo las sanciones establecidas en las leyes, la reproducción parcial o total de esta obra por cualquier medio o procedimiento, incluida la reprografía y el tratamiento informático, así como la distribución de ejemplares mediante alquiler o préstamo público.

ISBN: 978-84-16511-05-1
Depósito Legal: Z 70-2016
Impreso en España por Talleres Editoriales Cometa, S.A.

TEELL EDITORIAL: Todo Está En Los Libros

INTRODUCCIÓN

Ante la publicación en castellano de la trascendental obra de Nick Bostrom, *Superinteligencia*, se hace necesario recordar las palabras del gran filósofo español José Ortega y Gasset: "Uno de los temas que en los próximos años se va a debatir con mayor brío es el del sentido, ventajas, daños y límites de la técnica." Estas proféticas palabras de principios del siglo XX no han hecho más que cumplirse, siendo el presente libro un verdadero hito en la reflexión sobre la técnica o tecnología. Justo después de las citadas palabras de *Meditación de la técnica*, Ortega expresaba que "la misión del escritor es prever con holgada anticipación lo que va a ser problema, años más tarde, para sus lectores y proporcionarles a tiempo, es decir, antes de que el debate surja, ideas claras sobre la cuestión." Eso es precisamente lo que Bostrom ha buscado con *Superinteligencia*: presentar de manera clara y accesible un tema complejo y decisivo como es el problema de la inteligencia artificial de nivel sobrehumano.

La preocupación por las máquinas inteligentes no es algo completamente nuevo. Desde la segunda mitad del siglo pasado se ha convertido en una parte central de nuestro imaginario colectivo, con numerosas novelas y películas dedicadas al tema. En *Superinteligencia* veremos cómo se discuten, desde una base científica sólida, temas como la posibilidad de un gobierno mundial automatizado (recordando a clásicos como *Un mundo feliz*, de Aldous Huxley o *1984*, de George Orwell); la plausible catástrofe para la humanidad derivada del mal funcionamiento de las máquinas (un tema recurrente en el cine, como atestiguan *2001 Odisea en el espacio*, de Stanley Kubrick, *Terminator*, de James Cameron o *Matrix*, de los hermanos Wachowski); o las posibilidades y limitaciones de IAs con valores humanos (reflejadas en obras como *Yo, robot*, de Isaac Asimov o en la reciente *Her*, de Spike Jonze).

Las creaciones artísticas anticipan el mundo que está por venir, aunque en el caso de la IA los avances científico-técnicos suceden de una manera tan vertiginosa que la imaginación humana tiene dificultades para mantener el ritmo. El gran mérito de *Superinteligencia* es precisamente abordar de manera seria y rigurosa problemas que hasta hace no mucho parecían meras ocurrencias de ciencia-ficción. El libro de Bostrom es el mayor intento habido hasta la fecha de afrontar estos problemas desde una perspectiva estrictamente filosófica y científica, poniendo en juego un conocimiento realmente enciclopédico sobre la materia; un esfuerzo que sólo un pensador de la talla de Bostrom podía llevar a cabo.

En este sentido, es importante hacer notar que Bostrom tampoco juega a ser adivino, como sí hacen algunos gurús de la tecnología actuales. Bostrom no categoriza sobre la llegada de la IA ni sobre sus consecuencias. Sin embargo, la fuerza de sus argumentos nos sobresalta y nos despierta de nuestro letargo, haciéndonos ver el in-

menso peligro que traería consigo la llegada de una superinteligencia artificial. Esta toma de conciencia ni siquiera requiere por nuestra parte una gran fe en el progreso tecnológico; de hecho, el propio Bostrom se muestra escéptico respecto de las predicciones más optimistas sobre IA. Basta con atisbar el más que probable desastre al que abocaría un mal desarrollo de la IA para que nos preocupemos muy seriamente por estas tecnologías, confiemos plenamente en su desarrollo exitoso o no.

A pesar de lo que muchas veces se ha pensado —principalmente por su inclusión en la corriente del transhumanismo—, Nick Bostrom no es un tecnólatra. Tampoco un tecnófobo, aunque alguna lectura apresurada de *Superinteligencia* podría sacar esta conclusión. Bostrom es simplemente un pensador. Y como tal, su misión es reflexionar de manera radical sobre los problemas prominentes de su época. Su acierto consiste en haber identificado a la tecnología moderna como problema fundamental, y más concretamente a la fe ciega y el optimismo exacerbado que muchas personas profesan en la actualidad por la tecnología.

La obra lleva por subtítulo *Caminos, peligros, estrategias*, pero de estos términos acaba emergiendo uno por encima de los demás: el peligro. Necesitamos comprender los caminos que llevan a la superinteligencia para plantear las mejores estrategias frente a ella; pero la cuestión central en torno a la que giran todas las demás es el peligro que la superinteligencia trae consigo, la amenaza que proyecta sobre la humanidad y que deberíamos tener muy en cuenta. Este tono de advertencia no desemboca, en el caso de Bostrom, en un miedo paralizante que demonice todo avance en IA; únicamente nos hace comprender la enorme responsabilidad que tenemos y la imprudencia que sería desentenderse de estos problemas o confiar ingenuamente en su desarrollo benéfico.

A continuación sobrevolaremos brevemente el libro para dar una pequeña guía de lectura así como para aprovechar para comentar algunos puntos y términos especialmente problemáticos.

Superinteligencia se divide básicamente en dos bloques: un primer grupo de capítulos dedicado a presentar el problema de la inteligencia artificial, atendiendo a su historia, sus posibilidades presentes y su previsible desarrollo en los próximos años; y un segundo bloque algo más extenso dedicado a pensar qué podremos hacer con la superinteligencia, los peligros y problemas que probablemente acarree, y las estrategias y soluciones a nuestro alcance.

El primer bloque se extiende aproximadamente hasta el capítulo 6. Tras un breve recorrido en el primer capítulo a la historia de la IA y una somera exposición del estado de la cuestión, Bostrom pasa, ya en el segundo capítulo, a exponer los posibles caminos que podrían llevar hasta la superinteligencia. Este capítulo es crucial en tanto que introduce algunos términos omnipresentes a lo largo del libro, principalmente los de IA (inteligencia artificial) y ECC (emulación de cerebro completo). Sobre este punto conviene advertir del uso que Bostrom hace de inteligencia artificial en un sentido amplio para referirse a cualquier forma no humana de inteligencia, y un uso más restringido de inteligencia artificial para referirse a los tipos de IA cuyo origen no está directamente relacionado con el cerebro humano, frente a los que sí tienen dicho órgano como base (como la emulación de cerebro completo).

En el capítulo 3, Bostrom considera brevemente los distintos tipos de superinteligencia que podrían surgir, algo que retomará en más profundidad en el capítulo 6, cuando hable de los diferentes superpoderes cognitivos que una superinteligencia

presumiblemente tendría a su disposición. Bostrom introduce aquí por primera el decisivo concepto de Unidad, un ente único que concentraría todo el poder (político, tecnológico, económico, etc.) en sí mismo. Esta idea estará presente a lo largo de toda la obra y aparecerá repetidamente en numerosos párrafos.

Los capítulos 4 y 5 intentan atisbar la manera concreta en que la superinteligencia surgirá. Bostrom insiste en la relevancia del momento en que tenga lugar la explosión de inteligencia y en la importancia de su velocidad de despegue: cuanto antes suceda menos preparados estaremos. Bostrom propondrá una fórmula para calcular esta velocidad basada en la potencia de optimización y su contrapartida la resistencia al progreso (una construcción que traduce *recalcitrance*, término inexistente en castellano y cuyos sinónimos directos son inadecuados). La clave de estas discusiones y lo que más interesa al autor es la posibilidad de que la superinteligencia adquiera una ventaja estratégica decisiva, otro concepto que sobrevolará toda la obra.

Si bien, como ya hemos dicho, *Superinteligencia* es un libro unitario y muy bien trabado, el capítulo 7 marca la entrada en lo que podríamos considerar la segunda parte del libro. Este capítulo aborda por primera vez el problema de las relaciones entre inteligencia y motivación, un marco conceptual que servirá de base al resto de la obra. Un primer gran peligro en este sentido es el de antropomorfizar la IA, asumiendo que cuanto más inteligente sea la IA, más humana será y sus motivaciones y objetivos más se parecerán a los humanos. Esto podría no ser así, y, de hecho, si no nos esforzamos específicamente por conseguirlo, lo más probable es que la superinteligencia que surja sea profundamente inhumana.

El octavo capítulo se pregunta, de manera directa y sin rodeos, si estamos abocados al desastre. Este tono de preocupación gobernará la segunda parte del libro, en la que Bostrom tratará de hacerse cargo de las posibles amenazas de la superinteligencia, proponiendo, a su vez, las que considera mejores estrategias para afrontarlas, principalmente la prevención. Como hemos dicho, esta prudencia no debe confundirse con una tendencia al catastrofismo por parte del autor; la lectura de *Superinteligencia* no permite sacar esa conclusión. Lo que encontramos es una fría y objetiva mirada a la realidad, en la que se descubren unos riesgos que deben ser tenidos en cuenta. En este capítulo 8, Bostrom presenta diversos modos concretos en que la IA podría fallar, como la suplantación perversa (una IA que interpretara defectuosamente nuestras órdenes, suplantándolas por otras con efectos perjudiciales) o el crimen mental (la posibilidad de que las IAs o las emulaciones tengan un estatus moral que podría ser violado).

Los capítulos 9 y 10 están dedicados a examinar una primera salida a los problemas de la superinteligencia: los métodos de control. Éstos se dividen, por un lado, en métodos de control de la capacidad, que buscan impedir que la superinteligencia tenga un poder efectivo total sobre el mundo; y, por otro lado, en métodos de selección de la motivación, que buscan elegir los objetivos que la superinteligencia llegaría a tener. Bostrom demuestra que ambos métodos tiene carencias. En el caso del control de capacidad el problema está en la tensión existente entre minar las capacidades de la IA lo suficiente para que no tenga un poder absoluto, sin restarle tanta capacidad que ya no pueda ser considerada superinteligente. En el capítulo 10, Bostrom comparará diversos tipos de IA en función de su idoneidad para el control. La selección de la motivación es un problema más complejo, que por ello se trata más detalladamente en los últimos capítulos, donde encontramos la verdadera clave de la propuesta de Bostrom.

e la motivación con-
sobre la posibilidad
a que, en caso de ser
superinteligente. Por
parte de la obra, que
ero no cabe duda de
gías como de filoso-
os capítulos, pero sí
erente extrapolada),
propia capacidad de
perfecta los valores

tégico que probable-
va. Esto es estudiado
consecuencias socia-
a, para luego compa-
rinteligencias coexis-
todos los poderes se
donde el autor valora
o establecimiento de
y los acoplamientos
situaciones, solucio-
a de introducción de
las cuestiones funda-
cienciación sobre los
nos nuestros mejores

bir un libro accesible
de la traducción. La
os intentado transmi-
os, así como algunos
a mantenido el inglés
inos habituales en la
tilizado en referencia
tructuras inorgánicas
nado de Unidad (tra-
que diera a entender
a obra sirven también

eligencia es una gran
encia de su ejecución,
nguaje cercano y una
destinada a ser leída,
hasta cierto punto, el

ONSO FERNÁNDEZ

La inacabada fábula de los gorriones

Era la temporada de construcción de nidos, pero después de días de largo y duro trabajo, los gorriones se sentaron a la luz del atardecer, relajándose y trinando.

"Somos todos tan pequeños y débiles. ¡Imaginad lo fácil que sería la vida si tuviéramos una lechuza que nos ayudara a construir nuestros nidos!"

"¡Sí!", dijo otro. "Y podría ayudarnos a cuidar de nuestros ancianos e hijos".

"Podría darnos consejo y vigilar a los gatos que merodean cerca de aquí", añadió un tercero.

Entonces Pastus, el pájaro anciano y sabio, habló: "Enviemos exploradores en todas las direcciones e intentemos encontrar una lechuza joven abandonada en alguna parte, o quizás un huevo. Una cría de cuervo también serviría, o una comadreja muy joven. Podría ser lo mejor que jamás nos hubiera sucedido, al menos desde la apertura del «pabellón de grano ilimitado» del patio trasero".

La multitud estaba entusiasmada, y por todas partes los gorriones comenzaron a trinar con toda la fuerza de sus pulmones.

Sólo Scronkfinkle, un gorrión con un solo ojo y de temperamento inquisitivo, no estaba convencido de la sensatez de la empresa. Y dijo: "Eso seguramente sea nuestra perdición. ¿No deberíamos pensar antes en la labor de domesticar y controlar a la lechuza, antes de meter una criatura así entre nosotros?"

Pastus contestó: "Domesticar una lechuza parece una cosa muy difícil. Ya será bastante complicado encontrar un huevo de lechuza. Así que empecemos con eso. Cuando logremos criar a una lechuza, nos pondremos a pensar en emprender esa otra tarea".

"¡El plan tiene un defecto!", pio Scronkfinkle; pero sus protestas fueron en vano, pues la multitud ya había alzado el vuelo para empezar a llevar a cabo el plan ideado por Pastus.

Sólo dos o tres gorriones se quedaron rezagados. Juntos empezaron a intentar adivinar cómo se podría domesticar y controlar a una lechuza. Pronto se dieron cuenta de que Pastus tenía razón: era un reto extremadamente difícil, sobre todo por carecer de una lechuza real con la que practicar. No obstante, continuaron haciéndolo lo mejor que pudieron, con el temor constante a que la multitud regresara con un huevo de lechuza antes de encontrar una solución al problema de cómo controlar esa criatura.

No se sabe cómo termina la historia, pero el autor del presente libro lo dedica a Scronfinkle y sus seguidores.

, tiene
Es de-
psición
as más
entaja
mple-
ya que

inteli-
podría
pende
specie

remos
egiera
prác-
perin-
endre-
a poco
estaría

la su-
te, del
nunca.
o reto

avan-
n gra-
lad de
d. Los
rrollo
spués.
de la
ligen-
en el
ar las

multipolares 159
y hombres 160
rios y el desempleo 160
l y el bienestar 161
pio malthusiano desde una perspectiva histórica 163
nto demográfico e inversión 164
na economía algorítmica 166
id voluntaria, muerte ocasional 167
ijo máximamente eficiente sería divertido? 170
tratados inconscientes? 172
ción no es necesariamente hacia adelante 174
post-transición de una Unidad? 177
unda transición 177
ganismos y economías de escala 178
ión por tratado 180

o valores 185
de la introducción de valores 185
olutiva 187
por refuerzo 188
n de valores por asociación 189
notivacional 191
de valores 192
de emulaciones 201
tucional 202
207

s criterios para elegir 209
l de normatividad indirecta 209
herente extrapolada 211
explicaciones 212
para la VCE 213
servaciones 216
moralidad 217
quiero que hagas 220
pponentes 221
do de los objetivos 222
e la decisión 223
ología 224
ción 225
os lo suficiente 227

a estratégico 229
ientífica y tecnológica 229
llo tecnológico diferencial 230
cia en el orden de llegada 231

BLAS

o mundial (GDP). 3
tificial de nivel humano. 21
27
imágenes
31
to. 34
rección ortográfica
41
63
70
ncia. 77
rte de una IA. 96
a una hipotética
101
ena. 106
al del cerebro? 244
la IA. 248

12
ial de nivel humano? 20
la superinteligencia? 21
rebro completo. 32
re un conjunto
37
ferentes escenarios. 40
mente significativas. 81
relevantes y su conjunto
94
137
143

hibir una
ás rápido
ecimien-
lo peso a
o el "cre-
título de
:ia artifi-
r último,
os y reco-
avances.

aban col-
o incluso
ro común
erguida,
ente me-
varon un
s pueden
lar cultu-
quier otra

gías pro-
dos emi-
le la agri-
población
ores den-
acilidad y
pecializa-
uctividad
ados con

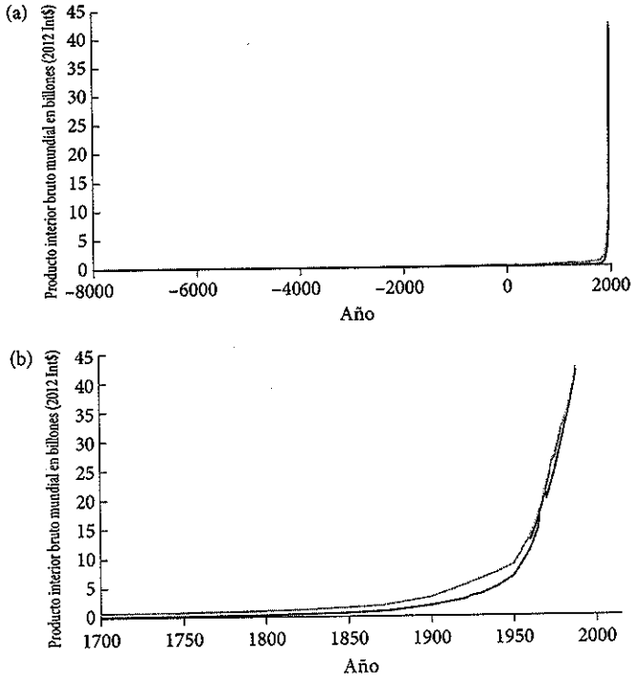
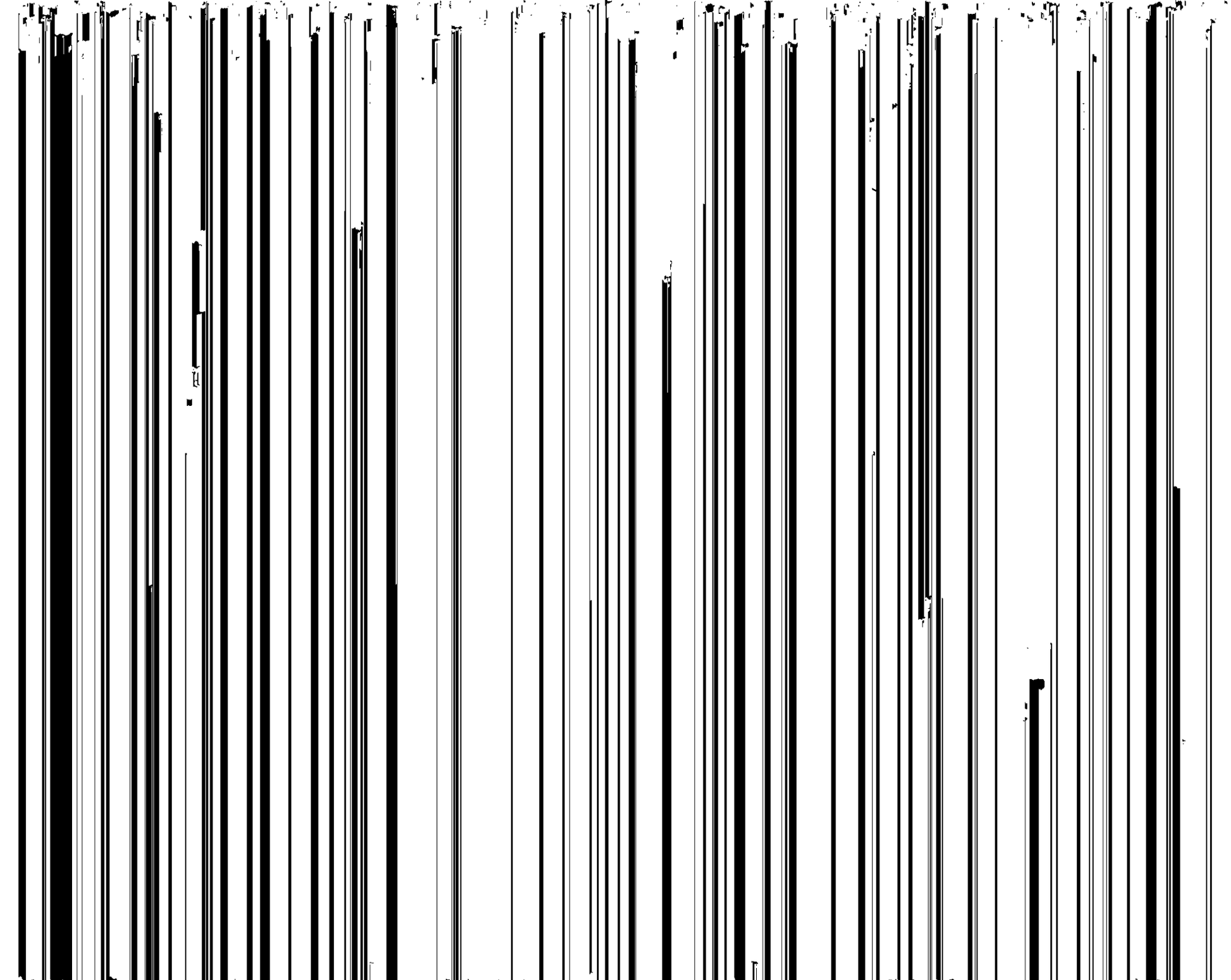


Figura 1. Historia a largo plazo del producto interior bruto mundial (GDP). Representada en una escala lineal, la historia de la economía mundial se asemeja a una línea pegada al eje X, hasta que súbitamente se inclina verticalmente hacia arriba. (a) Incluso cuando hacemos zoom a los últimos 10.000 años, el patrón sigue siendo esencialmente el de un ángulo de 90 grados. (b) Sólo en los últimos cien años más o menos la curva se separa perceptiblemente del nivel de cero. (Las distintas líneas del diagrama corresponden a distintas series de datos, que arrojan estimaciones ligeramente diferentes[6]).

el que el tiempo de duplicación de la economía mundial se reduzca a meras semanas que no conlleve la creación de mentes que sean mucho más rápidas y eficientes que las de tipo biológico familiar. Sin embargo, el hecho de tomarse en serio la perspectiva de una revolución en inteligencia artificial no tiene por qué depender de ejercicios de ajustes de curvas, ni de extrapolaciones a partir de antiguos crecimientos económicos. Como veremos, hay razones más convincentes para tomársela en serio.

Grandes expectativas

Que las máquinas se equiparen a los seres humanos en inteligencia general —es decir, en tener sentido común y una capacidad eficiente para aprender, razonar y planificar para afrontar complejos retos de procesamiento de información a lo largo de un amplio rango de ámbitos naturales y abstractos— ha sido algo esperado desde la invención de los ordenadores, en la década de los 40 del siglo XX. En aquella época, la llegada de este tipo de máquinas a menudo solía preverse para dentro de veinte años.[7]



de procesamiento. A
e toma de conciencia
de IA nunca podrían
er "invierno de la IA":
fondos y aumentó el

), cuando Japón lanzó
ón, una organización
os anteriores desarro-
que serviría de plata-
álgido de fascinación
en el que los líderes
samente descubrir la
car las mismas recetas
des en IA, otros países

de *sistemas expertos*.
, los sistemas expertos
encias a partir del co-
ertos humanos en ese
mal. Se construyeron
más pequeños conlle-
de desarrollar, validar
práctico adquirir un
grama. A finales de la
mbién se había extin-

tivos, lo mismo que le
ndo invierno de la IA
n toda justicia, que "la
ia, había consistido en
ente de fracasos al in-
iciales parecían apun-
on a huir de cualquier
entre los académicos
ado.[22]
en la década de los 90
legando a su deshielo.
técnicas, que parecían
do denominado "inte-
hioned Artificial Inte-
ipulación de símbolos
pertos de la década de
euronales y algoritmos
digma GOFAI, en par-
icos de IA (que solían
aunque sólo fuera una

n procedimiento eficiente para codificar soluciones
o que coincida con la estructura latente del ámbito
la evolución tiende a deambular eternamente en un
eda atrapada en un punto local óptimo. Aunque se
epresentación, la evolución es computacionalmente
r la explosión combinatoria.

goritmos genéticos son ejemplos de métodos que
década de los 90 del siglo XX, al surgir para ofre-
OFAI, que se había estancado. Pero no es nuestra
e estos dos métodos, ni elevarlos por encima de las
izaje artificial. De hecho, uno de los principales de-
veinte años ha sido una comprensión más clara de
diferentes pueden entenderse como casos especiales
común. Por ejemplo, muchos tipos de redes neu-
derarse clasificadores que realizan un tipo particu-
ación de máxima probabilidad).[26] Esta perspectiva
ompararse con una clase más amplia de algoritmos
r de ejemplos: "árboles de decisiones", "modelos de
máquinas de apoyo", "clasificadores bayesianos in-
os más cercanos", entre otros.[27] De manera similar,
len considerar la realización de una escalada esto-
onjunto de una clase más amplia de algoritmos de
algoritmos para la construcción de clasificadores, o
e soluciones, tiene su propio perfil de puntos fuertes
r matemáticamente. Los algoritmos difieren en los
npo de proceso y espacio de memoria, en los sesgos
a facilidad con que se puede incorporar contenido
mo de transparentes son sus procesos internos para

alrededor del aprendizaje artificial y de la resolu-
ntramos un conjunto de compensaciones matemá-
deal es el de un agente bayesiano perfecto, uno que
óptimo de la información disponible. Este ideal es
onalmente, es demasiado exigente para que sea im-
dor físico (véase el cuadro 1). En consecuencia, se
artificial como una búsqueda por encontrar atajos:
e flexiblemente al ideal bayesiano sacrificando algu-
les, a la vez que se preserva lo suficiente como para
los ámbitos de auténtico interés.

puede ver en el trabajo realizado durante el último
gráficos probabilísticos, como las redes bayesianas.
n una forma concisa de representar las relaciones de
condicionales que comparten algún ámbito particu-
ndependencia es esencial para superar la explosión
oblema tanto para la inferencia probabilística como
én nos ayudan de manera importante a entender el

a "noción de
ización se ha
de la ciencia,
e así, porque
ntento de ha-
xplicación de
subconjunto
e ordenador
a uno de los
o de mundos
r superan en
universo ob-
sibles en este
alizar cálculos

n físicamente
es heurísticas,
ín caso espe-
para agentes

mbito especí-
os algoritmos
nmediatas en
Monte Carlo,
y la genética
tes disciplinas
dísticas baye-
chos campos,
ca, la optimi-
ad de los pro-
de resultados
plicaciones de
enadores más

s ámbitos. La
ra que las IAs

	Sobrehumano en colaboración con humanos[42]	gana el campeonato estadounidense de Traveller (un futurista juego de guerra naval), lo cual lleva a cambiar las reglas para evitar sus estrategias poco ortodoxas.[43] Eurisko utilizaba la heurística para diseñar su flota, y también utilizaba la heurística para modificar su heurística.
Go	Nivel aficionado muy fuerte	En 2012, la serie Zen de programas dedicados a jugar al go ha alcanzado un rango de sexto dan en partidas rápidas (el nivel de un jugador aficionado muy fuerte) utilizando la búsqueda de árbol de variantes de Monte Carlo y técnicas de aprendizaje artificial.[54] Los programas que juegan al go han mejorado a un ritmo de más o

Sin embargo, hay un sentido importante en el que la IA de ajedrez fue un triunfo menor del que muchos esperaban. Antes se suponía, tal vez no sin razón, que para que un ordenador jugase al ajedrez a nivel de gran maestro, tendría que estar dotado de un alto grado de inteligencia *general*.[57] Se podía pensar, por ejemplo, que el gran juego del ajedrez requeriría ser capaz de aprender conceptos abstractos, pensar inteligentemente sobre estrategia, elaborar planes flexibles, realizar una amplia gama de ingeniosas deducciones lógicas y tal vez incluso de representarse el pensamiento del rival. No fue así. Resultó posible construir un programa de ajedrez perfectamente adecuado en torno a un algoritmo de objetivo específico.[58] Cuando se implementa en los rápidos procesadores disponibles a finales del siglo XX, da lugar a un nivel de juego muy fuerte. Pero una IA construida de esta manera es una cosa limitada. [illegible] al ajedrez, pero no puede hacer otra cosa."

En otros ámbitos, las soluciones han resultado ser *más* complicadas de lo previsto inicialmente, y el progreso más lento. El científico computacional Donald Knuth se sorprendió de que "la IA haya tenido ya éxito haciendo básicamente todo lo que requiere «pensar», pero haya fracasado en lo que la mayoría de la gente y los animales hacen «sin pensar» —eso, de algún modo, ¡es mucho más difícil!".[60] Analizar escenas visuales, reconocer objetos o controlar el comportamiento de un robot que interactúa con un entorno natural ha demostrado ser todo un reto. Sin embargo, se ha conseguido una buena cantidad de progreso y se sigue progresando, ayudado por constantes progresos en el hardware.

El sentido común y la comprensión del lenguaje natural también han resultado ser difíciles. En la actualidad, se suele pensar que alcanzar un rendimiento completamente humano en estas tareas es un problema de "IA completo", queriendo con esto decir [illegible] hacer, o estaría muy cerca de una capacidad general de esa clase.[62]

La maestría en ajedrez resultó ser alcanzable por medio de un algoritmo sorprendentemente simple. Resulta tentador especular que otras capacidades —tales como la capacidad de razonamiento general o alguna habilidad clave implicada en programación— podrían ser igualmente alcanzables a través de algún algoritmo sorprendentemente simple. El hecho de que se alcance un mejor rendimiento mediante un mecanismo complicado no significa que no exista un mecanismo simple que pueda hacer el trabajo igual o mejor. Podría suceder tan sólo que nadie hubiese encontrado esa alternativa más simple. El sistema de Ptolomeo (con la Tierra en el centro y el Sol, la Luna, los planetas y las estrellas orbitando en torno a ella) representó el estado de la técnica en astronomía durante más de mil años, y su exactitud predictiva fue mejorando con el paso de los siglos mediante una gradual complicación del modelo: añadiendo epiciclos sobre epiciclos a los movimientos celestes postulados. Después, todo el sistema fue derrocado por la teoría heliocéntrica de Copérnico, que era más simple, y —aunque sólo después de una mayor elaboración por parte de Kepler— más precisa en lo referente a las predicciones.[63]

tud de las aplicaciones. Aparte de las IA dedicadas a juegos reflejadas en la tabla 1, hay audífonos con algoritmos que filtran el ruido ambiental; buscadores de rutas que muestran mapas y ofrecen consejos de navegación a los conductores; sistemas de recomendación que sugieren libros y álbumes de música basándose en las compras y clasificaciones anteriores de un usuario; y sistemas de asistencia para decisiones médicas que ayudan a los médicos a diagnosticar el cáncer de mama, que recomiendan planes de tratamiento, y que ayudan en la interpretación de electrocardiogramas. Hay mascotas robóticas y robots de limpieza, robots que cortan el césped, robots de rescate, robots quirúrgicos y más de un millón de robots industriales.[64] La población mundial de robots supera los 10 millones.[65]

delos ocultos de Markov, se ha vuelto lo suficientemente preciso para su uso práctico (algunos fragmentos de este libro se elaboraron con la ayuda de un programa de reconocimiento de voz). Asistentes digitales personales, tales como Siri de Apple, responden a órdenes verbales y pueden responder preguntas sencillas y ejecutar órdenes. El reconocimiento óptico de caracteres de texto manuscrito y mecanografiado se utiliza rutinariamente en aplicaciones como la clasificación del correo y la digitalización de documentos antiguos.[66]

La traducción automática sigue siendo imperfecta pero es suficientemente buena para muchas aplicaciones. Los primeros sistemas utilizaban el enfoque GOFAI, con gramáticas de codificación manual que tuvieron que ser desarrolladas desde cero por lingüistas cualificados en cada idioma. Los nuevos sistemas utilizan técnicas estadísticas de aprendizaje artificial que construyen automáticamente modelos estadísticos a [illegible]

[illegible]

za en los pasos fronterizos automatizados de Europa y Australia. El Departamento de Estado de los Estados Unidos cuenta con un sistema de reconocimiento facial de más de 75 millones de fotografías para la tramitación de visados. Los sistemas de vigilancia utilizan una IA cada vez más sofisticada, además de tecnología de análisis de datos para analizar la voz, vídeos o textos, un buen número de los cuales se obtienen de los medios de comunicación electrónicos de todo el mundo y se almacenan en enormes bases de datos.

La demostración de teoremas y la resolución de ecuaciones están actualmente tan asentadas que apenas se siguen considerando como IA. En los programas informáticos científicos se incluyen solucionadores de ecuaciones como Mathematica. Los métodos de verificación formal, incluyendo demostradores de teoremas automatizados, suelen ser utilizados por fabricantes de chips de forma rutinaria para verificar el comportamiento de los diseños de circuitos antes de su producción.

Los organismos militares y de inteligencia de los Estados Unidos han liderado la senda que ha llevado al despliegue a gran escala de robots desactivadores de bombas, drones de vigilancia y ataque, y de otros vehículos no tripulados. Éstos todavía depen-

La programación inteligente es un ámbito muy importante en el que se ha tenido bastante éxito. La herramienta DART para la planificación y programación de logística automatizada se utilizó en la Operación Tormenta del Desierto, en 1991, hasta el punto de que la DARPA (la Agencia de Proyectos de Investigación Avanzada en Defensa de los Estados Unidos) afirma que esta aplicación, por sí sola, ha compensado sobradamente su inversión por treinta años en IA.[68] Los sistemas de reserva de las aerolíneas utilizan sofisticados sistemas de programación y de establecimiento de precios. Las empresas hacen un amplio uso de técnicas de IA en sus sistemas de control de inventario. También utilizan sistemas automáticos de reserva telefónica y líneas de ayuda asociadas a programas de reconocimiento de voz para guiar a sus desesperados clientes por laberintos de opciones de menú interconectadas.

Numerosos servicios de internet se basan en tecnologías de IA. Hay software encargado de vigilar el tráfico de correo electrónico por todo el mundo, y a pesar de la continua adaptación por parte de quienes envían spam para burlar las contramedidas tomadas contra ellos, los filtros de spam bayesianos han logrado mantener a raya, hasta cierto punto, la marea de correo basura. Softwares que utilizan componentes de IA son los responsables de aprobar o rechazar automáticamente las transacciones con tarjetas de crédito y vigilar continuamente la actividad de las cuentas para detectar los indicios de uso fraudulento. Los sistemas de recuperación de información también hacen un amplio uso del aprendizaje artificial. El motor de búsqueda de Google es, probablemente, el mayor sistema de IA que se ha construido hasta la fecha.

Ahora bien, hay que insistir en que la diferencia entre la inteligencia artificial y

mencionadas anteriormente podrían más bien considerarse aplicaciones genéricas que IA específicamente —aunque esto nos lleve de nuevo a la máxima de McCarthy de que, cuando algo funciona, ya no se le llama IA. Una distinción más relevante para nuestros propósitos es la que existe entre los sistemas que tienen un nivel reducido de capacidad cognitiva (se autodenominen "IA" o no) y los sistemas que tienen capacidad de resolución de problemas de aplicación más generalizada. Prácticamente, todos los sistemas actuales en uso son del primer tipo: de nivel reducido. No obstante, muchos de ellos contienen componentes que también podrían desempeñar un papel en la inteligencia artificial general del futuro, o estar al servicio de su desarrollo —componentes como los clasificadores, los algoritmos de búsqueda, los planificadores, los solucionadores de problemas y las infraestructuras de representación.

Un entorno de alto riesgo y muy competitivo en el que operan actualmente los sis-

hacen un uso generalizado de sistemas automatizados de intercambio de acciones. Aunque algunos de éstos son sólo formas de automatizar la ejecución de determinadas órdenes de compra o venta emitidas por un gestor de fondos humano, otros conllevan complejas estrategias comerciales que se adaptan a las condiciones cambiantes del mercado. Los sistemas analíticos utilizan diversas técnicas de análisis de datos y series temporales, con el objetivo de buscar patrones y tendencias en los mercados de valores, o para correlacionar los movimientos de precios históricos con variables externas, como palabras clave en los teletipos de los noticieros. Algunos proveedores de noticias financieras venden sistemas de transmisión de noticias que están especial-

Cuadro 2. El Flash Crash de 2010

La tarde del 6 de mayo de 2010, los mercados de acciones de Estados Unidos ya habían caído un 4% por las preocupaciones sobre la crisis de la deuda europea. A las 2:32 de la tarde, un gran vendedor (un complejo de fondos de inversión) inició un algoritmo de venta para deshacerse de un gran número de los futuros contratos de E-mini S&P

500 vendiéndolos a un ritmo vinculado a una medida de la actividad y la liquidez de la bolsa. Estos contratos fueron comprados por comerciantes algorítmicos de alta frecuencia, que fueron programados para eliminar rápidamente sus posiciones temporales a largo plazo mediante la venta de los contratos a otros comerciantes. Con la demanda de los principales compradores tocando fondo, los operadores algorítmicos comenzaron a vender E-Minis principalmente a otros operadores algorítmicos, que a su vez eran pasados a otros operadores algorítmicos, creando un efecto de "patata caliente" que elevó el volumen del comercio —lo cual fue interpretado por el algoritmo de venta como un indicador de alta liquidez, lo que le llevó a aumentar la velocidad a la que ponía contratos E-Mini en el mercado, acrecentando la espiral descendente. En algún momento, los operadores de alta frecuencia comenzaron a retirarse del mercado, secando la liquidez mientras que los precios seguían cayendo. A las 2:45 de la tarde, la negociación con E-Mini fue detenida por un interruptor automático, parando la funcionalidad lógica de la bolsa. Cuando se reinició el comercio, tan sólo cinco segundos más tarde, los precios se estabilizaron y pronto comenzaron a recuperar la mayor parte de las pérdidas. Pero

rrados del mercado, y los efectos secundarios habían llevado a un número sustancial de las operaciones de valores individuales a ejecutarse a precios "absurdos", como un centavo o 100.000 dólares. Después del cierre del mercado de ese día, los representantes de las bolsas se reunieron con los reguladores y decidieron cancelar todos las transacciones que habían sido ejecutadas a precios desviados en un 60% o más de sus niveles anteriores a la crisis (considerando este tipo de transacciones "claramente erróneas" y, por lo tanto, sujetas a cancelación post facto bajo las reglas comerciales existentes).[70]

Volver a contar aquí este episodio es una digresión, porque los programas informáticos implicados en el Flash Crash no eran especialmente inteligentes ni sofisticados, y el tipo de amenaza que crearon es fundamentalmente distinta de las preocupaciones que plantearemos más adelante en este libro en relación con la perspectiva de la superinteligencia artificial. En cualquier caso, estos acontecimientos ilustran varias lecciones útiles

ples (como el algoritmo de venta y los programas de negociación algorítmica de alta frecuencia) pueden producir efectos complicados e inesperados. El riesgo sistémico puede aumentar en un sistema a medida que se introducen nuevos elementos, unos riesgos que no son evidentes hasta después de que algo sale mal (y a veces ni siquiera entonces).[71]

Otra lección es que los profesionales sensatos podrían dar instrucciones a un programa basándose en hipótesis aparentemente sensatas y normalmente sólidas (por ejemplo, que el volumen del comercio es una buena medida de la liquidez del mercado), pero que esto puede producir resultados catastróficos cuando el programa continúa actuando siguiendo esas instrucciones con una férrea coherencia lógica, incluso en situaciones no

CAPÍTULO 2

Caminos hacia la superinteligencia

Actualmente las máquinas son muy inferiores a los humanos en inteligencia general. No obstante, algún día (como hemos sugerido) serán superinteligentes. ¿Cómo pasaremos de la situación actual a la situación prevista? Este capítulo explorará varios posibles caminos que la tecnología podría tomar. Abordaremos temas como la inteligencia artificial, la estimulación cerebral, la cognición biológica, y los interfaces hombre-máquina, así como las redes y organizaciones. Evaluaremos qué grado de plausibilidad tienen como caminos hacia la superinteligencia. La existencia de múltiples caminos aumenta la probabilidad de que la meta pueda alcanzarse al menos a través de uno de ellos.

Podemos definir tentativamente una superinteligencia como *cualquier intelecto que exceda en gran medida el desempeño cognitivo de los humanos en prácticamente todas*

[illegible] entre las diferentes formas de superinteligencia posibles. Por por ahora [illegible] Turing concibió un proceso repetitivo para desarrollar tal madurez mental.

experiencia consciente subjetiva puede ser muy importante para algunas cuestiones (en particular para cuestiones morales), nuestro interés ahora se centrará en los antecedentes causales y las consecuencias de la superinteligencia, no en teorías metafísicas de la mente[2].

Bajo esta definición, el programa de ajedrez Deep Fritz no es una superinteligencia, pues Fritz sólo puede considerarse inteligente en el reducido ámbito del ajedrez. Sin embargo, algunos tipos de inteligencia de ámbito reducido podrían ser relevantes. Cuando nos refiramos a un desempeño superinteligente que se limita a un ámbito particular, remarcaremos explícitamente esa restricción. Por ejemplo, una "superinteligencia ingenieril" sería un intelecto que supera enormemente a las mejores mentes

Inteligencia artificial

Los lectores de este capítulo no deben esperar unas instrucciones detalladas de cómo programar inteligencia artificial general. Dichas instrucciones no existen todavía, por supuesto. Y si estuviera en posesión de tales instrucciones, ciertamente no las publicaría en un libro. (Si las razones para esto no son inmediatamente obvias, los argumentos de capítulos subsiguientes lo aclararán).

Sí podemos, en cualquier caso, discernir algunas características generales del tipo de sistema que sería necesario. A estas alturas parece claro que la capacidad para aprender tendría que ser una característica integral del diseño básico de un sistema pensado para alcanzar la inteligencia general, no algo añadido al final como una extensión u ocurrencia de última hora. Lo mismo ocurre con la habilidad para manejarse frente a la incertidumbre y la información probable. Alguna capacidad para extraer conceptos útiles de los datos sensibles y estados internos, y para utilizar estos conceptos adquiridos dentro de representaciones combinativas usadas en el razonamiento lógico e intuitivo, también deberían formar parte de las características del diseño básico de una IA moderna pensada para lograr inteligencia general.

Los primeros sistemas de inteligencia artificial a la antigua usanza no se centraron, en su mayoría, en aprender, en manejar incertidumbre o en formar conceptos, quizás porque las técnicas que se ocupan de estas dimensiones estaban poco desarrolladas en aquel momento. Esto no significa que las ideas subyacentes sean completamente nuevas. La idea de usar el aprendizaje como lanzadera capaz de impulsar un sistema simple hasta el nivel de inteligencia humano se remonta al menos hasta el concepto de "máquina infantil" de Alan Turing, sobre la cual escribió en 1950:

> En lugar de tratar de producir un programa que simule la mente adulta, ¿por qué no tratar en su lugar de producir uno que simule la de un niño? Si la sometiéramos entonces a una educación apro-
>
> otra y comprobar si es mejor o peor. Hay una obvia conexión entre este proceso y la evolución… Podemos esperar, no obstante, que este proceso sea más rápido que la evolución. La supervivencia del más apto es un método lento de medir ventajas. El experimentador, gracias a su inteligencia, debería ser capaz de acelerar el proceso. Igualmente importante es el hecho de que él no estará restringido a mutaciones azarosas. Si puede identificar la causa de alguna debilidad, probablemente podrá idear la mutación que subsanará el problema[4].

Sabemos que procesos evolutivos no-dirigidos pueden producir niveles humanos de inteligencia general, puesto que ya lo han conseguido al menos una vez. Procesos evolutivos dirigidos —esto es, programas genéticos diseñados y guiados por un pro-

producir inteligencia, y llegar a la conclusión de que la ingeniería humana es ya muy superior a la evolución en algunas áreas y que, probablemente, llegue a ser superior en todas las áreas muy pronto. El hecho de que la evolución haya producido inteligencia indica, por lo tanto, que la ingeniería humana llegará pronto a conseguirlo también. Así, Moravec escribió (ya en 1976):

> La existencia de múltiples ejemplos de inteligencia diseñada bajo estas limitaciones debe darnos mucha confianza en que nosotros podamos conseguir lo mismo en poco tiempo. La situación es análoga a la historia del vuelo en seres más pesados que el aire, en la cual pájaros, murciélagos e insectos demostraron claramente la posibilidad de hacerlo antes de que la cultura lo dominara[6].

Debemos ser precavidos, no obstante, con las inferencias obtenidas mediante estos razonamientos. Es cierto que la evolución produjo vuelo en seres más pesados que el aire, y que ingenieros humanos tuvieron éxito en la misma empresa (si bien por medios de muy diferente naturaleza). Otros ejemplos podrían aducirse, como el sonar, la navegación magnética, las armas químicas, los fotorreceptores, y todo tipo de características mecánicas y cinéticas. Sin embargo, también podríamos fijarnos en las áreas en las que los ingenieros humanos han fracasado igualando a la evolución: en morfogénesis, en auto-reparación y en defensa inmunológica, por ejemplo, los esfuerzos humanos quedan muy lejos de los logros de la naturaleza. El argumento de Moravec, por tanto, no puede darnos "mucha confianza" en que alcanzaremos una inteligencia de nivel humano "muy pronto". En el mejor de los casos, la evolución de vida inteligente se sitúa en un nivel superior dentro de la intrínseca dificultad que conlleva diseñar inteligencia. Pero este nivel superior podría estar bastante por encima de las capacidades actuales de la ingeniería humana.

Otra manera de desplegar un argumento evolutivo en defensa de la posibilidad de la IA es a través de la idea de que podríamos, mediante algoritmos genéticos pro-

método específico mediante el cual la inteligencia podría ser producida.

Pero ¿es cierto que pronto tendremos suficiente potencia computacional para recapitular los procesos evolutivos relevantes que produjeron la inteligencia humana? La respuesta depende de lo que la tecnología computacional avance en las próximas décadas, y de cuánta potencia computacional se necesitará para procesar algoritmos genéticos con la misma capacidad de optimización que el proceso evolutivo de selección natural que reposa en nuestro pasado. Aunque, al final, la conclusión que obtenemos al seguir esta línea de razonamiento es decepcionantemente indeterminado, es productivo lanzar una estimación aproximada (véase cuadro 3). Aunque no consigamos otra cosa, este ejercicio nos permite llamar la atención sobre algunas incógnitas interesantes.

La conclusión es que los recursos computacionales requeridos simplemente para

Es plausible, no obstante, que comparado con la replicación por fuerza bruta de los procesos de evolución natural, se ganará mucha eficiencia diseñando el proceso de

Cuadro 3. ¿Qué se necesitaría para recapitular la evolución?

No todo logro alcanzado por la evolución en el curso del desarrollo de la inteligencia humana es relevante para un ingeniero humano que trate de hacer evolucionar la inteligencia en máquinas de manera artificial. Sólo una pequeña parte de la selección evolutiva que ha tenido lugar sobre la Tierra ha ido destinada a potenciar la inteligencia. Más específicamente, los problemas que los ingenieros humanos no son capaces superar de manera trivial puede que sean una porción muy pequeña del conjunto de la selección evolutiva. Por ejemplo, desde que nuestros ordenadores funcionan con energía eléctrica, no nos es necesario reconstruir las moléculas de la economía energética celular para crear máquinas inteligentes —si bien, la evolución molecular de tipo metabólico puede haber centrado los esfuerzos de una buena parte de la capacidad selectiva de la evolución a lo largo de la historia de la Tierra[7].

Uno puede argumentar que las ideas clave para la IA están contenidas en la estructura de los sistemas nerviosos, que surgieron hace menos de mil millones de años[8]. Si adoptamos esa perspectiva, entonces el número de "experimentos" relevantes disponibles para la evolución se reduce drásticamente. Existen en torno a $4\text{-}6\times10^{30}$, pero sólo 10^{19} insectos, y menos de 10^{10} humanos (aunque las poblaciones pre-agrarias tenían menor magnitud[9]). Estos números sólo son moderadamente apabullantes.

Los algoritmos evolutivos, no obstante, necesitan no sólo variaciones de entre las que seleccionar, sino una función de aptitud para evaluar las variantes, y esto suele ser el componente computacional más complicado. Una función de aptitud para la evaluación de la inteligencia artificial posiblemente necesitaría la simulación del desarrollo neural, del aprendizaje, y de la cognición para evaluar la aptitud. Por eso haríamos bien en no fijarnos en el mero número de organismos con sistemas nerviosos complejos, sino, en

estimación de esta última cuantía centrándonos en los insectos, que dominan la biomasa animal terrestre (sólo las hormigas componen el 15-20%[10]). El tamaño del cerebro de los insectos varía substancialmente, con insectos grandes y sociales que tienen grandes cerebros: un cerebro de abeja tiene poco menos de 10^6 neuronas, un cerebro de mosca de la fruta tiene 10^5, y el de las hormigas está en un lugar intermedio con 250.000 neuronas[11]. La mayoría de insectos pequeños tiene cerebros de sólo unos pocos miles de neuronas. Tirando por lo alto, si asignáramos a todos los 10^{19} de insectos un número de neuronas como el de la mosca de la fruta, el total sería de 10^{24} neuronas de insecto en el mundo. Este número podría aumentarse casi un orden mayor de magnitud para tomar en cuenta a los copépodos acuáticos, pájaros, reptiles, mamíferos, etc., para llegar a 10^{25}. (Por contraste, en los tiempos pre-agrarios había menos de 10^7 humanos, con menos de 10^{11} neuronas cada uno: con lo que habría menos de 10^{18} neuronas humanas en total, si

tiene la simulación. Modelos extremadamente simples de neuronas requieren en torno a 1.000 operaciones de coma flotante (FLOPS) por segundo para simular una neurona (en tiempo real). El modelo electrofisiológicamente realista de Hodgkin-Huxley utiliza

Cuadro 3. *Continúa*

1.200.000 FLOPS. Un modelo multi-compartimental más detallado añadiría entre tres o cuatro órdenes más de magnitud, mientras que modelos de mayor nivel que abstrajeran sistemas de neuronas podrían rebajar entre dos y tres órdenes de magnitud a partir de los modelos más simples[12]. Si quisiéramos simular 10^{25} neuronas a lo largo de más de mil millones de años de evolución (más tiempo del que llevan existiendo los sistemas nerviosos tal y como los conocemos), y dejáramos funcionar a nuestros ordenadores durante un año, tendríamos como resultado un requerimiento de entre 10^{31}-10^{44} FLOPS. En comparación, el Tianhe-2 de China, la supercomputadora más potente del mundo a septiembre de 2013, proporciona sólo 3.39×10^{16} FLOPS. En décadas recientes, han sido necesarios aproximadamente 6,7 años para que los ordenadores domésticos aumentaran

suficiente para llenar este vacío. Contar con hardware más especializado, o hacer funcionar los ordenadores más tiempo sólo aumentaría unos pocos órdenes de magnitud[15].

Esta cifra es prudente en otro sentido. La evolución ha alcanzado la inteligencia humana sin proponérsela como objetivo. En otras palabras, las funciones de aptitud para organismos naturales no selecciona sólo en función de la inteligencia y sus precursores[13]. Incluso entornos en los que organismos con habilidades superiores de procesamiento de información se vean recompensados pueden no seleccionar la inteligencia, porque las mejoras en inteligencia pueden (y suelen) imponer costes significativos, tales como un mayor consumo de energía o tiempos más lentos de maduración, y esos costes pueden imponerse a los beneficios obtenidos por un comportamiento más inteligente. Entornos muy hostiles también reducen el valor de la inteligencia: cuanto más corta sea la

niente de una habilidad progresiva para aprender. Una reducción de la presión selectiva por la inteligencia hace que se ralentice la expansión de innovaciones mejoradoras de la inteligencia, y así la oportunidad para que la selección favorezca subsiguientes innovaciones que dependen de ella. Y no sólo eso, sino que la evolución podría estancarse en un lugar óptimo que los humanos reconocerían y sobrepasarían mediante la alteración del equilibrio entre explotación y exploración o a través de una progresión de tests de inteligencia[14] gradualmente más difíciles. Y, como se ha mencionado antes, la evolución disemina mucho de su poder selectivo en rasgos que no están relacionados con la inteligencia (como las carreras de evolución competitiva al estilo Red Queen entre sistemas inmunes y parásitos). La evolución continúa gastando recursos produciendo mutaciones que pueden haberse mostrado consistentemente letales, y fracasa en aprovechar mutaciones distintas con efectos estadísticamente similares. Todo esto son ineficiencias de la selección natural (desde la perspectiva de quien intenta evolucionar la inteligencia) que serían relativamente fáciles de evitar para un ingeniero humano que usara algoritmos desarrolladores de software inteligente.

Es plausible que eliminar ineficiencias como las descritas redujeran varios órdenes de magnitud de los 10^{31}-10^{44} FLOPS calculados anteriormente. Desafortunadamente, es difícil saber exactamente cuántos órdenes de magnitud. Es difícil incluso hacer una estimación aproximada —pues, por lo que sabemos, las ganancias en eficiencia podrían ser de cinco órdenes de magnitud, o diez, o veinticinco.

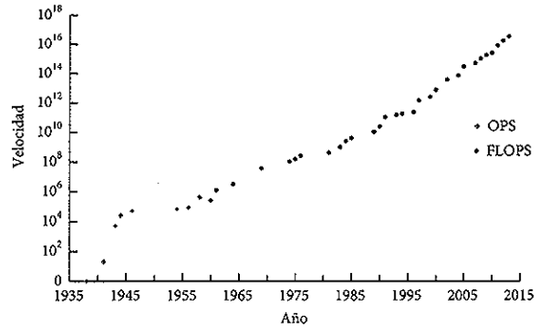


Figura 3. Rendimiento de las supercomputadoras. En un sentido limitado, "La Ley de Moore" se refiere a la observación de que el número de transistores en circuitos integrados se han doblado durante varias décadas

más general de que muchas métricas de desempeño usadas en tecnología computacional han seguido una tendencia exponencial similar. Aquí trazamos el pico de velocidad de la supercomputadora más rápida del mundo como una función temporal (en una escala logarítmica vertical). En años recientes, el crecimiento de la velocidad de serie de los procesadores se ha estancado, pero el aumento en el uso de la paralelización ha posibilitado que el número total de computaciones realizadas se mantenga en la tendencia[16].

en eficiencia. No podemos ni siquiera decir si supondrán cinco o veinticinco órdenes de magnitud. En ausencia de ulteriores elaboraciones, los argumentos evolutivos son incapaces de reducir significativamente nuestras expectativas sobre la dificultad de construir inteligencia humana en una máquina, ni pueden hablarnos sobre la escala

Hay otra complicación con este tipo de consideraciones evolutivas, una que hace que sea difícil derivar de ellas siquiera una estimación muy por lo alto de la dificultad de conseguir inteligencia evolutivamente. Debemos evitar el error de inferir del hecho de que vida inteligente evolucionó una vez en la tierra, que los procesos evolutivos implicados tuvieron una probabilidad razonablemente alta de producir inteligencia en un principio. Tal inferencia no es sólida porque no toma en cuenta el efecto de selección observacional que todos los observadores inevitablemente tendrán al haber nacido en un planeta donde surgió vida inteligente, sin importar cómo de probable o improbable era esto para cualquier planeta. Supongamos, por ejemplo, que, junto a los efectos sistemáticos de la selección natural fue necesario una cantidad enorme de *afortunada coincidencia* para producir vida inteligente —tanto es así que la vida inteligente evoluciona sólo en un planeta de cada 10^3 planetas, donde sólo surgen simples imitaciones. En ese caso, cuando pongamos a funcionar nuestros algoritmos genéticos para replicar lo que la evolución natural hizo, nos podríamos encontrar con que necesitaríamos realizar alrededor de 10^3 simulaciones antes de que encontremos una en la cual todos los elementos se organicen de la manera adecuada. Esto parece totalmente congruente con nuestra apreciación de que la vida sí evolucionó aquí en la Tierra. Sólo mediante razonamientos cuidadosos y algo intrincados —analizando instancias de evolución convergente de rasgos relativos a la inteligencia y enfrentándose a las sutilezas de la teoría de la selección observacional— podemos sortear parcialmente esta barrera epistemológica. A no ser que nos tomemos la molestia de hacer

esto, no estamos en posición de descartar la posibilidad de que la estimación por lo alto de los requerimientos computacionales necesarios para recapitular la evolución de la inteligencia expuesta en el Cuadro 3 sea demasiado baja por treinta órdenes de magnitud (o por otro número igualmente grande[17]).

Otra manera de argumentar a favor de la posibilidad de inteligencia artificial es fijándonos en el cerebro humano y sugiriendo que lo utilicemos como modelo para una inteligencia artificial. Podemos distinguir diferentes versiones de este enfoque basándonos en el grado en que se proponga imitar a las funciones cerebrales biológicas. En un extremo —uno en el que la imitación sería muy cercana— encontramos la idea de *emulación de cerebro completo*, que discutiremos en el siguiente apartado. En el otro extremo encontramos enfoques que se inspiran en el funcionamiento del cerebro pero no buscan una imitación a pequeña escala. Los avances en neurociencia y psicología cognitiva —que se verán ayudados por las mejoras de los instrumentales— deberían eventualmente desentrañar los principios generales del funcionamiento cerebral. Este conocimiento podría entonces guiar los esfuerzos en pos de la IA. Ya hemos mencionado las redes neuronales como un ejemplo de técnica de IA inspirada en el cerebro. La organización perceptual jerarquizada es otra idea trasvasada desde la ciencia cerebral al aprendizaje artificial. El estudio del aprendizaje por refuerzos ha sido motivado (al menos en parte) por su papel en las teorías psicológicas de la cognición animal, y las técnicas de aprendizaje por refuerzo (como por ejemplo el "Algoritmo-TD") inspiradas por estas teorías son ampliamente utilizadas en IA[18]. Con seguridad, más casos como éstos se irán sucediendo a lo largo del tiempo. Puesto que hay un número limitado —quizás un número muy pequeño— de mecanismos fundamentales distintos en el cerebro, un progreso continuo e incremental de la ciencia cerebral debería eventualmente descubrir todos estos mecanismos. Antes de que esto ocurra, sin embargo, es posible que un enfoque híbrido, combinando algunas técnicas inspiradas en el cerebro con algunos métodos puramente artificiales, alcance la meta. Es ese caso, el sistema resultante no tendría por qué ser reconocible como similar al cerebro, incluso aunque algunas claves derivadas del cerebro fueran utilizadas en su desarrollo.

La disponibilidad del cerebro como modelo proporciona un fuerte apoyo a la tesis de que la inteligencia artificial es factible en último término. Esto, no obstante, [illegible] porque es difícil predecir el ratio de

puede aplicarse igualmente a los enfoques neuromórficos y sintéticos de la inteligencia artificial.

Una variación del concepto de máquina infantil de Turing es la idea de "IA seminal[19]". Mientras que una máquina infantil, como Turing la imaginó, hubiera tenido una arquitectura fija que simplemente desarrollaría sus potencialidades inherentes por la *acumulación de contenido*, una IA seminal sería una inteligencia artificial más sofisticada capaz de mejorar su propia *arquitectura*. En las etapas tempranas de la IA seminal, tales mejoras podrían ocurrir principalmente por ensayo y error, adquisición de información o asistencia por parte de los programadores. En etapas posteriores, no obstante, una IA seminal debería ser capaz de *comprender* su funcionamiento lo suficiente como para producir nuevos algoritmos y estructuras computacionales que impulsaran su desempeño cognitivo. Esta necesaria comprensión podría ser el resultado de que la IA seminal alcanzara un nivel suficiente de inteligencia general en muchos ámbitos, o sobrevenir cuando se superara un cierto umbral en un ámbito particularmente relevante como podría ser la ciencia computacional o las matemáticas.

Esto nos lleva a otro importante concepto, el de "auto-mejoramiento recursivo". Una IA seminal exitosa sería capaz de mejorarse repetidamente a sí misma: una versión temprana de la IA podría diseñar una versión mejorada de sí misma, y la versión mejorada —siendo más inteligente que la original— podría ser capaz de diseñar una versión aún más inteligente de sí misma, y así sucesivamente[20]. Bajo estas circunstancias, tal proceso de auto-mejoramiento recursivo podría continuar lo suficiente como para resultar en una explosión de inteligencia —un evento en el que, en un breve período de tiempo, el nivel de inteligencia de un sistema aumentaría desde una dote modesta de capacidades cognitivas (quizás sub-humanas en la mayoría de sentidos, pero con un talento específico en codificación y búsqueda de IA) hasta la superinteligencia radical. Volveremos sobre esta importante posibilidad en el capítulo 4, donde las dinámicas de tal evento se analizarán pormenorizadamente. Nótese que este modelo indica la posibilidad de sorpresas: los intentos de construir inteligencia artificial general pueden fallar básicamente de plano hasta que el último y crítico factor sea incluido, momento en el cual una IA seminal puede llegar a ser capaz de realizar un auto-mejoramiento recursivo de manera sostenida.

Antes de terminar esta subsección, hay otra cosa que deberíamos enfatizar, y es [illegible] no tiene por qué parecerse a una mente humana. IAs

que cuanto más lejos en el futuro miremos, más posible será que [illegible] de la funcionalidad cerebral al descubierto hasta el punto de poder crear inteligencia artificial de esta manera.

Diferentes personas trabajando en pos de la inteligencia artificial mantienen diferentes visiones sobre cómo de prometedoras son las aproximaciones neuromórficas comparadas con aproximaciones que aboguen por diseños completamente sintéticos. La existencia de pájaros demostró que el vuelo de objetos más pesados que el aire era físicamente posible y propició los esfuerzos para construir máquinas voladoras. Pero los primeros aviones funcionales no batían sus alas. No podemos estar seguros de que la inteligencia de las máquinas será como el vuelo, que los humanos consiguieron mediante un mecanismo artificial, o como la combustión, que llegamos a dominar imitando la manera en que ocurren naturalmente los fuegos.

La idea de Turing de diseñar un programa que adquiera la mayoría de su conte-

Deberíamos esperar que tengan arquitecturas cognitivas muy [illegible] gencias biológicas, y en las primeras etapas de su desarrollo tendrán perfiles muy diferentes de fortalezas y debilidades (aunque, como argumentaremos más tarde, éstas podrían sobreponerse a cualquier debilidad inicial). Además, los objetivos sistémicos de las IAs podrían diferir radicalmente de los de los seres humanos. No hay motivo para esperar que una IA genérica esté motivada por el amor o el odio o el orgullo o algún otro típico sentimiento humano: estas complejas adaptaciones requerirían un deliberado y costoso esfuerzo para que se recrearan en las IAs. Esto es a la vez un gran problema y una gran oportunidad. Volveremos sobre la cuestión de la motivación de la IA en capítulos posteriores, pero es tan central al argumento de este libro que merece la pena mantenerlo en mente durante su recorrido.

Emulación de cerebro completo

En la emulación de cerebro completo (también conocida como "subida a la nube"), software inteligente sería producido escaneando y modelando minuciosamente la estructura computacional de un cerebro biológico. Este enfoque representa así un caso límite de inspiración en la naturaleza: el plagio directo. Alcanzar la emulación de cerebro completo requiere seguir los siguientes pasos.

Primero, crear un escaneado del cerebro suficientemente detallado. Esto puede implicar estabilizar el cerebro post-mortem a través de la vidrificación (un proceso que transforma el tejido en una especie de vidrio). Una máquina podría entonces diseccionar el tejido en láminas, que serían entregadas a otra máquina para que las escanearan, quizás por un conjunto de microscopios electrónicos. Diversos colorantes podrían ser aplicados en esta etapa para poner al descubierto diversas propiedades químicas y estructurales. Muchas máquinas escaneadoras podrían trabajar en paralelo para procesar múltiples láminas de cerebro simultáneamente.

Segundo, los datos en bruto provenientes de los escáneres se pasarían a un ordenador que construyera la red neuronal en tres dimensiones que implementó cognición al cerebro original. En la práctica, este paso puede hacerse a la vez que el primero para reducir la cantidad de imágenes en alta resolución almacenadas. El mapa resultante sería entonces combinado con una biblioteca de modelos neurocomputacionales de los diferentes tipos de neuronas o de diferentes elementos neuronales (como tipos particulares de conectores sinápticos). La Figura 4 muestra algunos resultados de los procesos de escaneado y procesamiento de imágenes producidos por tecnología actual.

En la tercera etapa, la estructura neurocomputacional resultante del paso previo es implementada en un ordenador lo suficientemente potente. Si es completamente exitoso, el resultado sería una reproducción digital del intelecto original, con la memoria y la personalidad intacta. La mente humana emulada ahora existiría como software en un ordenador. La mente podría habitar una realidad virtual o interactuar con el mundo exterior mediante apéndices robóticos.

El camino de la emulación de cerebro completo no nos exige que comprendamos cómo funciona la cognición humana ni cómo se programa una inteligencia artificial. Sólo requiere que entendamos características funcionales básicas de los elementos computacionales del cerebro. No es necesario un fundamental avance conceptual ni teórico para el éxito de la emulación de cerebro completo.

La emulación de cerebro completo requiere, sin embargo, de algunas tecnologías realmente avanzadas. Hay tres prerrequisitos clave: (1) *escaneado*: microscopios de alto rendimiento con suficiente resolución y capacidad para detectar propiedades relevantes; (2) *traducción*: análisis automatizado de imágenes que transforme los datos brutos en modelos interpretados tridimensionalmente de los componentes neuronales relevantes; y (3) *simulación*: hardware suficientemente potente como para implementar la estructura computacional resultante (véase tabla 4). (En comparación con estos difíciles pasos, la construcción de una realidad virtual básica o un cuerpo robótico con un canal audiovisual de recepción y algún canal de salida es relativamente fácil. Una estructura de recepción/salida simple pero mínimamente adecuada parece factible ya con la tecnología del presente[23]).

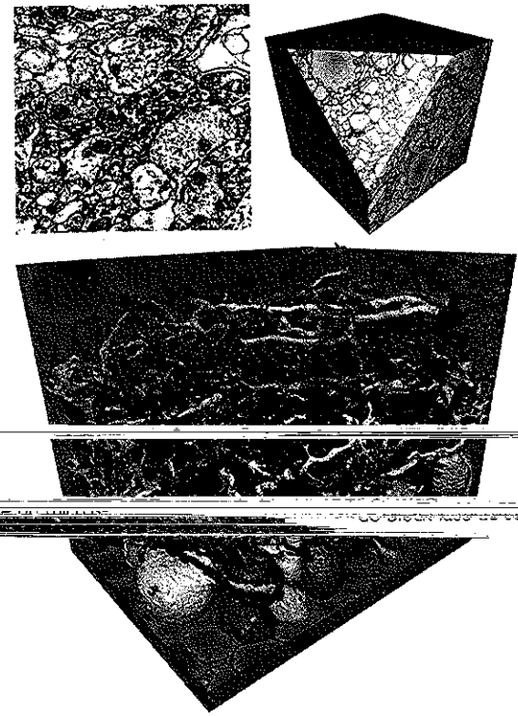


Figura 4. Reconstrucción 3D neuroanatómica a partir de imágenes de un microscopio electrónico. *Arriba a la izquierda:* una micrografía electrónica típica que muestra secciones transversales de materias neuronales —dendritas y axones—. *Arriba a la derecha:* imagen de volumen de tejido neural de la retina de conejo obtenida por escaneado electrónico. Imágenes individuales en 2D han sido apiladas en un cubo (con un lado de aproximadamente 11 micras). *Conclusión:* Reconstrucción de un subconjunto de las proyecciones neuronales llenando un volumen de neuropilo, generado por un algoritmo de segmentación automatizado[22].

Hay buenas razones para pensar que las tecnologías que harán posible lo anterior son alcanzables, aunque no en un futuro próximo. Ya existen modelos computacionales razonables de muchos tipos de neuronas y procesos neuronales. Se ha desarrollado software de reconocimiento de imagen que puede rastrear axones y dendritas a través de una pila de imágenes bidimensionales (aunque es necesario mejorar su fiabilidad). Y hay herramientas de visualización que proporcionan la resolución necesaria —con un microscopio de efecto túnel es posible "ver" los átomos individuales, en una resolución muy superior a la requerida. Sin embargo, a pesar de que los conocimientos y capacidades actuales parecen sugerir que no existe un impedimento intrínseco al desarrollo de las necesarias tecnologías posibilitantes, está claro que todavía se necesita una gran cantidad de progreso técnico para hacer posible la emulación de cerebro completo humano.[24] Por ejemplo, la tecnología de microscopía necesitaría no sólo suficiente resolución sino también suficiente rendimiento. El uso de un microscopio de escaneado de túnel de resolución atómica para mapear la superficie necesaria sería demasiado lento para ser realizable. Sería más plausible utilizar un microscopio electrónico de menor resolución, pero esto requeriría nuevos métodos para la preparación y tinción de tejido cortical que hicieran visibles detalles relevantes como la estructura sináptica fina. También sería necesaria una gran expansión de las bibliotecas neurocomputacionales y grandes mejoras en el procesamiento automatizado de imágenes e interpretación de escaneado.

En general, la emulación de cerebro completo se basa menos en el conocimiento teórico y más en la capacidad tecnológica que la inteligencia artificial. La cantidad de tecnología necesaria para la emulación cerebro completo depende del nivel de abs-

Tabla 4. Capacidades necesarias para la emulación de cerebro completo

Escaneado	Pre-procesamiento / fijación		Preparación adecuada de cerebros, conservando la microestructura pertinente y estado.
	Estado físico		Métodos de manipulación de cerebros fijos y piezas de tejido antes, durante y después del escaneado.
	Creación de imágenes	*Volumen*	Capacidad para escanear volúmenes de cerebro completos en un tiempo y precio razonables
		Resolución	Escanear a una resolución suficiente como para permitir la reconstrucción
		Información funcional	Habilidad del escaneado para identificar las propiedades funcionalmente relevantes del tejido
Traducción	Procesamiento de imágenes	*Ajuste geométrico*	Ajustar las distorsiones debidas a imperfecciones del escáner
		Interpolación de datos	Ajuste de datos omitidos
		Eliminación del ruido	Mejora de la calidad del escaneado
		Rastreo	Detectar estructuras y procesarla en un modelo 3D consistente del tejido
	Interpretación del escaneado	*Identificación del tipo de células*	Identificar el tipo de células
		Identificación de sinapsis	Identificar sinapsis y su conectividad
		Estimación de parámetros	Estimar parámetros funcionalmente relevantes de células, sinapsis y otras entidades
	Modelado del software del sistema neurológico	*Modelado matemático*	Modelo de las entidades y su comportamiento
		Implementación eficiente	Implementación del modelo
Simulación	Almacenamiento		Almacenamiento del modelo original y su estado actual
	Ancho de banda		Comunicación eficiente entre procesadores
	CPU		Potencia del procesador para hacer funcionar a la simulación
	Simulación corporal		Simulación de un cuerpo que posibilite una interacción con un entorno virtual o un entorno real como robot
	Simulación del entorno		Entorno virtual para un cuerpo virtual

tracción en que se emule el cerebro. En este sentido hay un equilibrio entre conocimiento y tecnología. En general, cuanto peor sea nuestro equipo de exploración y cuanto menos potentes sean nuestros ordenadores, menos podremos confiar en la simulación de procesos electrofisiológicos y químicos del cerebro, y más comprensión teórica de la arquitectura computacional que estamos tratando de emular será necesaria para crear representaciones abstractas de las funcionalidades relevantes.[25] Por el contrario, con una tecnología de escaneo suficientemente avanzada y abundante potencia de cálculo, podría ser posible forzar una emulación incluso con un conocimiento bastante limitado de cerebro. En un irreal caso límite, podríamos imaginar emular un cerebro al nivel de sus partículas elementales utilizando la ecuación de Schrödinger de la mecánica cuántica. En ese caso, se podría confiar enteramente en el conocimiento de la física existente, desechando cualquier modelo biológico. Este caso extremo, sin embargo, demandaría una potencia de cálculo y adquisición de datos imposible. Un nivel mucho más plausible de emulación sería uno que incorporara las neuronas individuales y su matriz de conectividad, junto con algunas neuronas de la estructura de sus árboles dendríticos y tal vez algunos estados variables de sinapsis individuales. Las moléculas neurotransmisoras no se simularían individualmente, sino que las concentraciones fluctuantes se modelarían de manera poco refinada.

Para evaluar la viabilidad de la emulación de cerebro completo hay que entender cuál es la clave para el éxito. El objetivo no es crear una simulación del cerebro tan detallada y precisa que uno podría utilizarla para predecir exactamente lo que habría ocurrido en el cerebro original si hubiera sido sometido a una secuencia particular de estímulos. El objetivo, en cambio, es captar lo suficiente las propiedades computacionalmente funcionales del cerebro para permitir que la emulación resultante pueda llevar a cabo trabajo intelectual. Para este fin, muchos de los caóticos detalles biológicos de un cerebro real son irrelevantes.

Un análisis más elaborado distinguiría entre los diferentes niveles de éxito de emulación basándose en la medida en que se hubiera conservado la funcionalidad de procesamiento de información del cerebro emulado. Por ejemplo, se podría distinguir

[illegible]

que fuera capaz de hacer el mismo trabajo intelectual que el cerebro emulado; y (3) *una emulación genérica* (que también podría estar distorsionada) que fuera algo así como un niño, que careciera de las habilidades o los recuerdos que habrían sido adquiridas por el cerebro adulto emulado, pero con la capacidad de aprender más de lo que un ser humano normal puede aprender.[26]

Aunque en última instancia parece factible producir una emulación de alta fidelidad, parece bastante probable que la primera emulación de cerebro completo que lograríamos si fuésemos por este camino sería de inferior nivel. Antes de poder conseguir que las cosas funcionaran perfectamente, probablemente se conseguiría que funcionaran de manera imperfecta. También es posible que un impulso hacia la tecnología de emulación diera lugar a la creación de algún tipo de IA neuromórfica que adaptara algunos principios neurocomputacionales descubiertos gracias a los esfuerzos de emulación y así hibridar éstos con métodos sintéticos, siendo posible que esto ocurriera antes de llevar a cabo una emulación de cerebro completo integral y funcional. La posibilidad de que convergiera en una IA neuromórfica de este tipo,

Curiosamente, el rendimiento decreciente se para en gran medida cuando la selección se extiende sobre varias generaciones. Por lo tanto, seleccionar repetidamente al número 1 de 10 durante más de diez generaciones (donde cada nueva generación se compone de los descendientes de los seleccionados en la generación anterior) producirá un mayor incremento en el valor de rasgo que una selección de 1 de 100 de una sola vez. El problema con la selección secuencial, por supuesto, es que se necesita más tiempo. Si cada paso generacional tarda veinte o treinta años, entonces tan sólo cinco generaciones sucesivas podrían llevarnos hasta bien entrado el siglo XXII. Mucho antes de esto, formas más directas y poderosas de ingeniería genética (por no hablar de inteligencia artificial) estarán probablemente disponibles. Hay, sin embargo, una tecnología complementaria, la cual, una vez que haya sido desarrollada para su uso en

persona más inteligente en la población humana a lo largo de la historia. Un mundo que tuviera una gran población de estos individuos podría (si tuviera la cultura, la educación, la infraestructura de comunicaciones, etc. que lo acompañara) constituir una superinteligencia colectiva.

El impacto de esta tecnología será empañado y retrasado por varios factores. Existe un retraso madurativo inevitable hasta que los embriones finalmente seleccionados se convierten en seres humanos adultos: por lo menos veinte años antes de que un niño mejorado llegue a una productividad total, más todavía antes de que estos niños lleguen a constituir un segmento importante de la población activa. Además, incluso después de que la tecnología se haya perfeccionado, las tasas de adopción probablemente comenzarán a la baja. Algunos países podrían

embargo, todavía persisten algunos retos científicos sustanciales, como la traducción de resultados en animales a seres humanos, o como la prevención de alteraciones epigenéticas en las líneas de células madre derivadas. Según un experto, estos retos pueden tener aplicaciones humanas a "10 o incluso 50 años vista."[47]

Con gametos derivados de células madre, la cantidad de poder selectivo disponible para una pareja podría aumentar en gran medida. En la práctica actual, un procedimiento de fertilización in vitro implica normalmente la creación de menos de diez embriones. Con gametos derivados de células madre, unas pocas células donadas pueden convertirse en un número virtualmente ilimitado de gametos, los cuales pueden ser combinados para producir embriones, que podrían entonces ser genotipados o secuenciados, eligiendo los más prometedores para la implantación. Dependiendo del coste de la preparación y selección de cada embrión individual, esta tecnología

que el niño tendría un gran talento y estaría libre de predisposiciones genéticas a la enfermedad—. Menores costes de atención médica y la expectativa de un sueldo vitalicio más alto también servirán de argumento a favor de la selección genética. Si el uso del procedimiento se volviera más común, sobre todo entre las élites sociales, puede haber un cambio cultural en las normas de crianza que hará ver el uso de la selección como algo que las parejas responsables y formadas hacen. Muchos de los inicialmente reacios podrían subirse al carro con el fin de tener un hijo que no está en desventaja respecto de los hijos mejorados de sus amigos y colegas. Algunos países podrían ofrecer incentivos para alentar a sus ciudadanos a tomar ventaja de la selección genética con el fin de aumentar la reserva de capital humano del país, o para aumentar la estabilidad social a largo plazo mediante la selección de rasgos como la docilidad, la obediencia, la sumisión, la conformidad, la aversión al riesgo,

que constara de los siguientes pasos:

1. Genotipado y selección de un número de embriones destacados en las características genéticas deseadas.
2. Extracción de células madre de esos embriones y conversión de los mismos en espermatozoides y óvulos, madurando dentro de seis meses o menos.[49]
3. Cruzar el nuevo esperma y óvulos para producir embriones.
4. Repetir hasta que se acumulen grandes cambios genéticos.

De esta manera, sería posible llevar a cabo diez o más generaciones de selección en pocos años. (El procedimiento consumiría mucho tiempo y sería caro; sin embargo, en principio, sólo sería necesario hacerlo una vez en lugar de repetirlo para cada

taria, en cierta medida, en competencia con otros atributos deseados, como la salud, la belleza, la personalidad, o el atletismo. La selección de embriones por iteración, al ofrecer una gran cantidad de poder selectivo, rebajaría esta competencia, lo que permitiría una fuerte selección simultánea para varios rasgos. Sin embargo, este procedimiento tiende a alterar la relación genética normal entre padres e hijos, algo que podría afectar negativamente a la demanda en muchas culturas.[51]

Con nuevos avances en tecnología genética, puede ser posible sintetizar genomas específicos, obviando la necesidad de grandes reservas de embriones. La síntesis de ADN ya es algo rutinario en biotecnología y está en gran medida automatizado, aunque todavía no es factible sintetizar un genoma humano entero que pudiera ser utilizado en un contexto reproductivo (en buena parte debido a las dificultades aún

es impactos de la selección genética en diferentes escenarios[52]

nología	"FIV +" Selección de 1 de 2 embriones [4 puntos]	"FIV agresivo" Selección de 1 de 10 embriones [12 puntos]	"Óvulo in vitro" Selección del 1 d embriones [19 pu	"La selección de embriones por iteración" [más de 100 puntos]
tilidad pción	Socialmente insignificante durante una generación. Efectos de controversia social más importante que los impactos directos.	Socialmente insignificante durante una generación. Efectos de controversia social más importante que los impactos directos.	Una mejora conti forma una minori en posiciones de selección cognitiv	Los seleccionados dominan la primera línea de las élites de científicos, abogados, médicos, ingenieros. ¿Renacimiento intelectual?
e" n	Impacto cognitivo leve en primera generación, se combina con la selección de rasgos no cognitivos para dar una ventaja perceptible a una minoría.	Gran parte de los estudiantes de Harvard son mejorados. La segunda generación domina profesiones cognitivamente exigentes.	Los seleccionados las filas de científi abogados, médicc ingenieros de pri generación.	"Posthumanidad"[53]
idad" ción	Discapacidad de aprendizaje mucho menos frecuente entre los niños. En la segunda generación, la población por encima de los umbrales de un coeficiente intelectual alto se ha duplicado.	Un crecimiento sustancial en el nivel de instrucción e ingresos. La segunda generación aumenta varias magnitudes en la dirección adecuada.	El CI bruto típico científicos emine veces más comú primera generaci de veces en la se generación.	"Posthumanidad"

están presentes de manera poco frecuente en la población, pero que pueden tener efectos positivos significativos en la cognición.[55]

Una intervención que se hará posible cuando los genomas humanos puedan sintetizarse es la prueba de "corrección ortográfica" genética de un embrión. (La selección de embriones por iteración también podría permitir una versión aproximada de esto). Cada uno de nosotros lleva actualmente una carga mutacional, de cientos de mutaciones quizás, que reducen la eficiencia de los diversos procesos celulares.[56] Cada mutación individual tiene un efecto casi imperceptible (que por ello va lentamente eliminándose del acervo genético); sin embargo, de manera combinada, tales mutaciones puede suponer una gran carga para nuestro funcionamiento.[57] Las diferencias indivi-

tes de las mutaciones acumuladas. En términos provocativos, se podría decir que los individuos creados a partir de estos genomas revisados podrían ser "más humanos" que nadie actualmente vivo, en el sentido de que serían expresiones menos distorsionadas de la forma humana. Tales personas no serían puras copias idénticas, porque los seres humanos varían genéticamente de otros modos que no implican a las mutaciones perjudiciales. Pero la manifestación fenotípica de un genoma revisado puede llegar a ser una constitución física y mental excepcional, con un elevado nivel de funcionamiento en dimensiones de rasgo poligénico como la inteligencia, la salud, la resistencia, y la apariencia.[58] (Una vaga analogía podría hacerse con las caras compuestas, en el que los defectos de los individuos superpuestos se equilibran: véase la Figura 6).

Otras técnicas biotecnológicas potenciales también podrían ser relevantes. La clonación humana reproductiva, una vez lograda, se podría utilizar para replicar el genoma de individuos excepcionalmente talentosos. La adopción de este sistema estaría limitada por la preferencia de la mayoría de los futuros padres por estar biológica-

Figura 6. Las caras compuestas como una metáfora de corrección ortográfica para genomas. Cada una de las imágenes centrales fue producida por la superposición de fotografías de dieciséis individuos diferentes (re-

mediante la eliminación de mutaciones individuales, corregir genomas puede producir gente más cercana

mente relacionados con sus hijos; pero la práctica podría, no obstante, llegar a tener un impacto relevante debido a que (1) incluso un aumento relativamente pequeño en el número de personas excepcionalmente talentosas podrían tener un efecto significativo; y (2) es posible que algún Estado se embarcara en un programa de eugenesia a gran escala, tal vez mediante el pago a madres de alquiler. Otros tipos de ingeniería genética —tales como el diseño de nuevos genes sintéticos o la inserción en el genoma de regiones promotoras y otros elementos de control de la expresión genética— también podrían llegar a ser importantes con el tiempo. Posibilidades aun más exóticas pueden darse, tales como cubas llenas de tejido cortical cultivado de estructura compleja, o animales transgénicos "mejorados" (tal vez algunos mamíferos de gran cerebro, como la ballena o el elefante, enriquecidos con genes humanos). Estos últimos casos son totalmente especulativos, pero en un período de tiempo más largo no pueden ser completamente descontados.

Hasta ahora hemos hablado de las intervenciones en la línea germinal, que se pueden hacer en los gametos o embriones. Mejoras somáticas de genes, sin pasar por el ciclo de la generación, podrían producir, en principio, impacto de manera más rápida. Sin embargo, tecnológicamente son mucho más difíciles. Requieren que los genes modificados puedan insertarse en un gran número de células del cuerpo vivo, incluyendo, en el caso de la mejora de la cognición, en el cerebro. La selección entre las células o embriones de óvulos existentes, en contraste, no requiere la inserción de genes. Incluso este tipo de terapias como la de la línea germinal no implican la modificación del genoma (por ejemplo, la corrección del genoma o el empalme en alelos raros) son mucho más fáciles de implementar en el gameto o la etapa embrionaria, donde se está tratando con un pequeño número de células. Por otra parte, las intervenciones en la línea germinal de embriones probablemente pueden lograr mayores efectos que las intervenciones somáticas en los adultos, debido a que las primeras serían capaces de dar forma al desarrollo temprano del cerebro, mientras que estas últimas estarían limitadas a trabajar sobre una estructura existente. (Algunas de las cosas que podrían hacerse a través de la terapia génica somática también podría ser alcanzables por medios farmacológicos).

Centrándonos, por tanto, en las intervenciones en la línea germinal, debemos tener en cuenta cómo el desfase generacional retrasaría cualquier gran impacto a nivel mundial.[60] Incluso si la tecnología se perfeccionara hoy y empezara a utilizarse de inmediato, se necesitarían más de dos décadas para que una progenie mejorada ge-

[illegible]

ticas de tratamiento de fertilidad estándar e información genética para elegir entre [illegible] embriones que de otro modo habrían sido seleccionados por casualidad.

Los retrasos también pueden ser el resultado de obstáculos no relacionados con un miedo al fracaso (con demandas de pruebas de seguridad), sino relacionados con el miedo al éxito, con la demanda de regulación impulsada por las preocupaciones sobre la legitimidad moral de la selección genética o sus implicaciones sociales más amplias. Tales preocupaciones son probablemente más influ[illegible]

Alemania de posguerra, por ejemplo, ha optado por dar un gran rodeo frente a las prácticas reproductivas que podrían ser percibidas, incluso en la forma más remota, como dirigidas a la mejora, una postura que es comprensible dada la historia particularmente oscura de atrocidades conectadas al movimiento eugenésico de ese país. Otros países occidentales tienden a adoptar un enfoque más liberal. Y algunos países —tal vez China o Singapur, los cuales tienen políticas de población a largo plazo— podrían no sólo permitir sino promover activamente el uso de la selección genética y la ingeniería genética para mejorar la inteligencia de sus poblaciones, una vez que la tecnología para hacerlo esté disponible.

Cuando el ejemplo se haya establecido, y los resultados empiecen a mostrarse, los reticentes tendrán fuertes incentivos para seguir dicho ejemplo. Las naciones se enfrentarían a la posibilidad de convertirse en remansos cognitivos y quedarse atrás en lo referente a científicos, militares, y concursos de prestigio económicos respecto de los competidores que adoptaran las nuevas tecnologías de mejora humana. Los individuos dentro de una sociedad verían cómo las plazas de las escuelas de élite se llenan de niños seleccionados genéticamente (que también pueden en promedio ser más guapos, más saludables y más conscientes) y querrán tener las mismas ventajas para sus propios hijos. Hay alguna posibilidad de que un gran cambio de actitud aconteciera en un tiempo relativamente corto, tal vez en tan sólo una década, una vez que se haya demostrado que la tecnología funciona y que proporciona un beneficio sustancial. Las encuestas de opinión en Estados Unidos muestran un cambio dramático en la aprobación pública de la fecundación in vitro después del nacimiento del primer "bebé probeta", Louise Brown, en 1978. Unos años antes, sólo el 18% de los estadounidenses dijo que usarían en su caso la FIV para tratar la infertilidad; sin embargo, en una encuesta realizada poco después del nacimiento de Louise Brown, el 53% dijo que lo haría, y el número ha seguido subiendo.[61] (En comparación, en una encuesta realizada en 2004, el 28% de los estadounidenses aprueba la selección de embriones para fomentar "la fuerza o la inteligencia", el 58% la aprueba para evitar el cáncer en adultos, y el 68% la aprueba para evitar enfermedades fatales en la infancia.[62])

Si sumamos los diversos retrasos —digamos de cinco a diez años para reunir la información necesaria para que la selección sea significativamente efectiva entre un conjunto de embriones fecundados in vitro (posiblemente mucho más tiempo antes de que los gametos derivados de células estén disponibles para su uso en la reproducción humana), diez años para construir una implantación importante, y de veinte a

[illegible]

cas. La velocidad de este mejoramiento se aceleraría en gran medida cuando grandes grupos de individuos concebidos mediante tecnologías genéticas más potentes y de nueva generación (en gametos derivados de células madre y en particular a través de la selección de embriones por iteración) entren a formar parte de la fuerza laboral.

Con el pleno desarrollo de las tecnologías genéticas descritas anteriormente (dejando a un lado las posibilidades más exóticas como la inteligencia en el tejido neural [illegible]

en comparación con nuestros parientes cercanos los grandes simios y con nuestros propios ancestros humanoides; y no hay ninguna razón para suponer que el *homo sapiens* haya alcanzado el vértice de la eficacia cognitiva alcanzable para un sistema biológico. Lejos de ser la especie biológica más inteligente posible, estamos probablemente mejor representados como la especie biológica más estúpida capaz de iniciar una civilización tecnológica —un nicho que llenamos porque llegamos a ese punto primero, no porque estamos en ningún sentido óptimamente adaptados a ello.

Progresar a lo largo de la ruta biológica es claramente factible. El desfase generacional de las intervenciones en la línea germinal significa que el progreso no puede ser tan repentino y abrupto como en escenarios que impliquen inteligencia artificial. (Las

parece poco probable que este tipo de interfaces puedan utilizarse ampliamente como mejoras en ningún momento cercano.[65]

Para empezar, existen riesgos significativos de complicaciones médicas —incluyendo infecciones, desplazamiento del electrodo, hemorragia, y declive cognitivo— cuando se implantan electrodos en el cerebro. Tal vez el ejemplo más vívido hasta la fecha de los beneficios que se pueden obtener a través de la estimulación cerebral es el tratamiento de pacientes con Parkinson. El implante para tratar el Parkinson es relativamente simple: no se comunica realmente con el cerebro sino simplemente suministra una corriente eléctrica estimulante para el núcleo subtalámico. Un vídeo de demostración muestra a un sujeto desplomado en una silla, completamente in-

puede obtener una idea de la magnitud de la brecha considerando la diferencia de velocidad entre los componentes electrónicos y las células nerviosas: incluso los transistores de hoy en día operan a una velocidad diez millones de veces menor que la de las neuronas biológicas). Sin embargo, incluso las mejoras relativamente moderadas de la cognición biológica podrían tener consecuencias importantes. En particular, la mejora cognitiva podría acelerar la ciencia y la tecnología, incluidos los avances hacia formas más potentes de amplificación de la inteligencia biológica y de la inteligencia artificial. Consideremos cómo el ritmo de avance en el campo de la inteligencia artificial podría cambiar en un mundo donde "hombre promedio Joe" es un par intelectual

cerebrales profundos mostró reducciones en la fluidez verbal, la atención selectiva, el nombramiento de colores, y la memoria verbal en comparación con los sujetos de control. Los sujetos tratados también informaron de otras quejas cognitivas.[66] Tales riesgos y efectos secundarios pueden ser tolerables si el procedimiento se utiliza para aliviar una discapacidad severa. Pero para que los sujetos sanos pidan neurocirugía para sí mismos, tendría que aspirarse a una mejora muy sustancial de la funcionalidad normal.

Esto nos lleva a la segunda razón para dudar de que la superinteligencia se logre

tres conclusiones: (1) por lo menos formas leves de superinteligencia son alcanzables por medio de mejoras biotecnológicas; (2) la viabilidad de humanos cognitivamente mejorados aumenta la probabilidad de que haya formas de inteligencia artificial factibles —porque incluso si *nosotros* fuéramos fundamentalmente incapaces de crear inteligencia artificial (que no hay ninguna razón para suponerlo así), la inteligencia artificial aún podría estar al alcance de humanos cognitivamente mejorados; y (3) si

retinas artificiales.[68] Pacientes con Parkinson o con dolor crónico podrían beneficiarse de la estimulación cerebral profunda que excita o inhibe la actividad en un área particular del cerebro.[69] Lo que parece mucho más difícil de lograr es una interacción directa de banda ancha entre cerebro y ordenador para proporcionar aumentos sustanciales en la inteligencia de una forma que no pudiera alcanzarse más fácilmente por otros medios. La mayoría de los potenciales beneficios que los implantes cerebra-

mucho por aumentar la velocidad a la que pensamos y aprendemos, a no ser que toda la maquinaria neuronal necesaria para dar sentido a los datos se actualizara de manera similar. Dado que esto incluiría a casi todo el cerebro, lo que realmente se necesitaría sería una "prótesis total de cerebro", lo cual es sólo otra forma de decir inteligencia artificial general. Sin embargo, si dispusiéramos de una IA con inteligencia de nivel humano, se podría prescindir de la neurocirugía: una computadora puede tener una carcasa de metal lo mismo que una de huesos. Así que este caso límite sólo nos lleva de vuelta a la senda de la IA, que ya hemos examinado.

La interfaz cerebro-ordenador también se ha propuesto como una manera de sa-

cursor en una pantalla mediante el pensamiento.[72] El ancho de banda alcanzado en tales experimentos es bajo: el paciente minuciosamente escribe una lenta letra tras otra a un ritmo de sólo unas palabras por minuto. Uno puede imaginar fácilmente versiones mejoradas de esta tecnología, tal vez un implante de última generación podría conectarse al área de Broca (una región en el lóbulo frontal que participa en la producción del lenguaje) y reproducir el discurso interno.[73] No obstante, aunque esta tecnología podría ayudar a algunas personas con discapacidades inducidas por accidente cerebrovascular o por degeneración muscular, sería poco atractiva para los sujetos sanos. La funcionalidad que proporcionaría es esencialmente la de un micrófono unido a un software de reconocimiento de voz, que ya está disponible comer-

tracraneal). Mantener nuestras máquinas fuera de nuestro cuerpo también hace más fácil el actualizarse.

Pero ¿qué pasa con el sueño de sobrepasar el lenguaje por palabras y establecer una conexión entre dos cerebros que permita "descargarse" conceptos de una mente a otra, pensamientos o áreas enteras de experiencia? Podemos descargar archivos de gran tamaño a nuestras computadoras, incluyendo bibliotecas con millones de libros y artículos, y esto se puede hacer en segundos: ¿podría hacerse algo similar con nuestro cerebro? La aparente plausibilidad de esta idea deriva probablemente de una visión incorrecta de cómo se almacena y se representa la información en el cerebro. Como se ha señalado, el obstáculo para alcanzar la inteligencia humana no es la rapidez en que los datos en bruto pueden introducirse en el cerebro, sino más bien lo rápido que el cerebro puede extraer significado y dar sentido a los datos. Tal vez se sugerirá que transmitimos significados directamente, en lugar de paquetes de datos sensoriales que deben ser decodificados por el receptor. Hay dos problemas con esto. La primera

datos estandarizados. Más bien, cada cerebro desarrolla sus propias representaciones idiosincrásicas de contenido de nivel superior. Qué conjuntos neuronales particulares

en las células separadas de memoria desplegadas claramente.[74] Por lo tanto, no sería posible establecer una correspondencia sencilla entre las neuronas de un cerebro y las de otro, de tal manera que los pensamientos pudieran deslizarse automáticamente de uno al otro. A fin de que los pensamientos de un cerebro sean inteligibles para otro, los pensamientos deben ser descompuestos y empaquetados en símbolos siguiendo alguna convención compartida que permita al cerebro de recepción interpretarlos correctamente. Ésta es la misión del lenguaje.

En principio, podríamos imaginar la descarga del trabajo cognitivo de articulación e interpretación en una interfaz que de algún modo fuera capaz de leer los estados

leer con habilidad y escribir simultáneamente tal vez miles de neuronas

de manera individual, crear la interfaz requerida es probablemente un problema de IA completo. La interfaz tendría que incluir un componente capaz de mapear (en tiempo real) patrones de encendido en un cerebro y convertirlos en patrones de encendido semánticamente equivalentes en otro cerebro. La comprensión detallada de múltiples niveles en computación neuronal necesaria para llevar a cabo tal tarea parecería permitir directamente una IA neuromórfica.

A pesar de estas reservas, la ruta ciborg hacia la mejora cognitiva no es del todo sombría. El impresionante trabajo sobre el hipocampo en ratas ha demostrado la viabilidad de una prótesis neural que puede mejorar el rendimiento en la memoria de

dos docenas de

en un número similar de neuronas de otro área ("CA1"). Un microprocesador está capacitado para discriminar entre dos patrones de activación diferentes en la primera zona (correspondiente a dos memorias diferentes, "palanca derecha" o "palanca izquierda") y para aprender cómo estos patrones se proyectan en la segunda zona. Esta prótesis no sólo es capaz de restaurar sus funciones cuando se bloquea la conexión neuronal normal entre las dos áreas neuronales, sino que mediante el envío de una señal especialmente clara de un patrón particular de la memoria a la segunda área, se puede mejorar el rendimiento de la memoria más allá de lo que la rata es normalmente capaz de hacer. Aunque sea una gran dificultad técnica para los estándares contemporáneos, el estudio deja muchas preguntas difíciles sin respuesta: ¿Hasta qué punto puede aplicarse este enfoque a un mayor número de memorias? ¿Hasta qué punto podemos controlar la explosión combinatoria, que de no ser controlada amenazaría con hacer que el aprendizaje de asignación correcta fuera imposible a medida que aumentara el número de neuronas de entrada y salida? ¿La mejora de rendimiento en la tarea

de la capacidad para desaprender la asociación

tos de prueba pueden todavía beneficiarse de alguna manera incluso si —a diferencia de las ratas— pudieran acogerse a las ayudas de memoria externas tales como el lápiz

CAPÍTULO 3

Formas de superinteligencia

Entonces, ¿qué entendemos exactamente por "superinteligencia"? Aunque no deseamos empantanarnos en discusiones terminológicas, hay que decir algo para aclarar el terreno conceptual. Este capítulo identifica tres formas diferentes de superinteligencia y argumenta que son, en un sentido relevante para la práctica, equivalentes. También muestra que hay más posibilidades de que la inteligencia se dé en un sustrato artificial que en un sustrato biológico. Las máquinas tienen una serie de ventajas fundamentales que les darán una superioridad abrumadora. Los humanos biológicos, aun mejorados, serán superados.

Muchas máquinas y animales no humanos ya se desempeñan a niveles sobrehumanos en ámbitos específicos. Los murciélagos interpretan señales de sónar mejor que el hombre, las calculadoras nos superan en aritmética, y los programas de ajedrez nos ganan al ajedrez. La gama de tareas específicas que un software realiza mejor continuará expandiéndose. Pero si bien los sistemas especializados de procesamiento de información tendrán muchos usos, hay profundas cuestiones adicionales que surgen sólo ante la perspectiva de intelectos artificiales que tuvieran suficiente inteligencia general como para sustituir a los seres humanos en todos los ámbitos.

Como se indicó anteriormente, utilizamos el término "superinteligencia" para referirnos a intelectos que superan considerablemente a las mejores mentes humanas actuales en muchos dominios cognitivos muy generales. Esto es muy vago todavía. Diferentes tipos de sistemas con atributos de rendimiento muy dispares podrían ser calificados como superinteligencias bajo esta definición. Para avanzar en el análisis, es útil desglosar esta simple noción de superinteligencia distinguiendo diferentes vertientes de supercapacidades intelectuales. Hay muchas maneras para hacer tal descomposición. Aquí vamos a diferenciar entre tres formas: la superinteligencia de velocidad, la superinteligencia colectiva y la superinteligencia de calidad.

Superinteligencia de velocidad

Una superinteligencia de velocidad es un intelecto igual a una mente humana, pero más rápido. Ésta es la forma más fácil de analizar conceptualmente una superinteligencia.[1] Podemos definir superinteligencia por velocidad de la siguiente manera:

Superinteligencia de velocidad: *un sistema que puede hacer todo lo que el intelecto humano puede hacer, pero mucho más rápido.*

Por "mucho" queremos decir algo así como "varios órdenes de magnitud." Pero en lugar de tratar de borrar todos los restos de vaguedad de la definición, vamos a confiar en que el lector haga una interpretación razonable.[2] El ejemplo más simple de superinteligencia de velocidad sería una emulación de cerebro completa que se ejecutara en hardware rápido.[3] Una emulación que funcionara a una velocidad diez mil veces más rápida que un cerebro biológico sería capaz de leer un libro en pocos segundos y escribir una tesis doctoral en una tarde. Con un factor de aceleración de un millón, una emulación podría lograr todo un milenio de trabajo intelectual en un día de trabajo.[4]

Para una mente tan rápida, los acontecimientos en el mundo exterior parecerían [illegible] silencio deslizándose por el espacio hacia una cita con un planeta lejano; contemplando cómo la anticipación del choque inminente se propaga lentamente a través de los pliegues de la materia gris de su amigo, y de allí hacia su sistema nervioso periférico; podría observar cómo su cuerpo gradualmente toma la forma de un "¡Uy!" congelado —en lo que le daría tiempo no sólo a pedir otra taza, sino también a leer un par de artículos científicos y echarse una siesta.

Debido a esta aparente dilatación del tiempo del mundo material, una superinteligencia de velocidad preferiría trabajar con objetos digitales. Podría vivir en la realidad virtual y hacer frente a dicha economía informacional. Alternativamente, podría interactuar con el medio físico por medio de operadores de nanoescala, ya que las extremidades en escalas tan pequeñas podrían funcionar más rápido que los apéndices macroscópicos. (La frecuencia característica de un sistema tiende a ser inversamente proporcional a su escala de longitud.[5]) Una mente rápida se comunicaría principalmente con otras mentes rápidas en lugar de con tardígrados seres humanos cuasi-vegetales.

La velocidad de la luz se convierte en un obstáculo cada vez más importante cuanto más rápidas se vuelvan las mentes, ya que las mentes más rápidas se enfrentarían a mayores costes de oportunidad a la hora de usar su tiempo para viajar o para comunicarse a larga distancia.[6] La luz es aproximadamente un millón de veces más rápida que un avión a reacción, por lo que se necesitaría un agente digital con una aceleración mental de 1.000.000 ×, aproximadamente, la misma cantidad de tiempo subjetivo que un viajero humano contemporáneo necesita para viajar por todo el mundo. Hacer una llamada de larga distancia a alguien tomaría tanto tiempo como llegar allí "en persona", aunque sería más barato porque una llamada requeriría menos ancho de banda. Los agentes con grandes aceleraciones mentales que quisieran conversar extensamente podrían encontrar ventajoso trasladarse unos cerca de otros. Las mentes extremadamente rápidas con necesidad de interacción frecuente (como los miembros de un equipo de trabajo) podrían establecer su residencia en computadoras ubicadas en el mismo edificio para evitar demoras frustrantes.

Superinteligencia colectiva

Otra forma de superinteligencia podría consistir en un sistema que lograra un rendi-

tres formas de superinteligencia es, por tanto, el mismo. En este sentido, el alcance indirecto de la inteligencia humana actual es similar, si suponemos que finalmente seremos capaces de crear algún tipo de superinteligencia. Sin embargo, hay un sentido en el que las tres formas de superinteligencia están mucho más cerca una de la otra: cualquiera de ellas podría crear otras formas de superinteligencia más rápidamente de lo que nosotros podríamos hacer desde nuestro estado actual.

Los *alcances directos* de las tres formas diferentes de superinteligencia son más difíciles de comparar. Puede que no haya un orden definitivo. Sus respectivas capa-

una superinteligencia de calidad, y así sucesivamente. [illegible] superinteligencia de velocidad destaca en tareas que requieren

mientras que la superinteligencia colectiva destaca en tareas que permiten la descomposición analítica en sub-tareas paralelas y en tareas que exigen la combinación de diferentes perspectivas y habilidades. En algún sentido vago, la superinteligencia de calidad sería la forma más capaz de todas, en la medida en que podría comprender y resolver problemas que están, en un sentido práctico, fuera del alcance *directo* de la superinteligencia de velocidad y de la superinteligencia colectiva.[14]

En algunos ámbitos, la cantidad es un mal sustituto de la calidad. Un genio solitario trabajando en una habitación descorchada puede escribir *En busca del tiempo perdido*. ¿Podría una obra maestra equivalente producirse mediante la contratación [illegible]?[15] Incluso dentro de la gama actual

conjuntos de muchos mediocres. Si ampliamos nuestro espectro para incluir en el [illegible] posibilidad de que haya problemas

intelectuales que sólo sean solucionables por una superinteligencia y [illegible] quibles para cualquier conjunto, del tamaño que sea, de seres humanos no mejorados.

Puede, por lo tanto, haber algunos problemas que fueran solucionables por una superinteligencia de calidad, y tal vez por una superinteligencia de velocidad, pero que una superinteligencia colectiva poco integrada no pudiera resolver (a no ser que primero amplificara su propia inteligencia).[16] No podemos ver claramente la naturaleza de todos estos problemas, pero podemos caracterizarlos en términos generales.[17] Éstos tienden a ser problemas que implican múltiples interdependencias complejas que no permiten soluciones independientemente verificables: problemas que, por lo tanto, no pueden ser resueltos de una manera gradual, y para los que podríamos re-

Fuentes de ventaja para la inteligencia digital

Pequeños cambios en el volumen cerebral y en el cableado pueden tener importantes consecuencias, como vemos cuando comparamos los logros intelectuales y tecnológicos de los seres humanos con los de otros simios. Los cambios mucho más grandes en recursos informáticos y arquitectura que la inteligencia artificial permitirá, probablemente tendrán consecuencias aún más profundas. Es difícil para nosotros, quizás imposible, imaginarnos en un sentido intuitivo las aptitudes de una superinteligencia;

un microprocesador moderno (~ 2 GHz).[19] Como consecuencia de ello, el cerebro humano se ve obligado a confiar en la paralelización masiva y es incapaz de llevar a cabo rápidamente cualquier cálculo que requiera un gran número de operaciones secuenciales.[20] (Cualquier cosa que el cerebro hace en menos de un segundo no puede requerir más de un centenar de operaciones secuenciales, quizás sólo unas pocas docenas). Sin embargo, muchos de los algoritmos más importantes en sentido práctico para la programación y para la informática no pueden paralelizarse de manera fácil. Muchas de las tareas cognitivas podrían realizarse mucho más eficientemente si el soporte nativo del cerebro para algoritmos paralelizables buscadores de patrones se complementara con, y estuviera integrado por ayudas para el

de 120 m/s o menos, mientras que los núcleos de procesamiento electrónicos pueden co-

[illegible] unidad mínima como de grande puede ser un cerebro biológico al mismo tiempo que funciona como una unidad de procesamiento individual. Por ejemplo, para conseguir un retardo de ida y vuelta de menos de 10 ms entre dos elementos de un sistema, los cerebros biológicos deben ser menores que 0,11 m^3. Un sistema electrónico, por otro lado, podría ser de $6{,}1 \times 10^{17}$ m^3, aproximadamente del tamaño de un planeta enano: dieciocho órdenes de magnitud más grande.[22]

- *Número de elementos computacionales.* El cerebro humano tiene un poco menos de 100 mil millones de neuronas.[23] Los seres humanos tienen un cerebro casi tres veces y media más grande que el de los chimpancés (aunque sólo una quinta parte del tamaño del cerebro de los cachalotes).[24] El número de neuronas en una criatura biológica está obviamente limitado

dad de RAM de una computadora digital, es evidente que las ventajas de hardware de las inteligencias digitales harían posible que éstas tuvieran memorias de trabajo más grandes. Esto podría permitirle a esas mentes entender intuitivamente relaciones complejas que los seres humanos sólo pueden manejar tentativamente a través de calculaciones perseverantes.[28] La

velocidad con la que se puede acceder sería, por tanto, mucho mayor en un cerebro artificial que en un cerebro biológico.

- *Fiabilidad, vida útil, sensores, etc.* Las inteligencias artificiales pueden tener varias otras ventajas de hardware. Por ejemplo, las neuronas biológicas son menos fiables que los transistores.[30] Mientras que la compleja computación requiere esquemas de codificación redundantes que utilizan múltiples elementos para codificar un solo bit de información, un cerebro digital podría conseguir algunas mejoras a través de la eficiencia en la utilización de elementos de computación fiables y de alta precisión. Los cerebros se fatigan después de unas horas de trabajo y comienzan a decaer de forma permanente después de unas décadas

cambia lentamente (aunque algunos detalles de la conectividad sináptica pueden cambiar en escalas de tiempo más cortas, como días).[31]

En la actualidad, el poder computacional del cerebro biológico todavía se compara favorablemente con el de las computadoras digitales, aunque la supercomputadoras punteras están alcanzando niveles de rendimiento que están dentro del rango de estimaciones plausibles de potencia de procesamiento del cerebro.[32] Pero el hardware está mejorando rápidamente, y los límites últimos de rendimiento para el hardware son muy superiores a los de los sustratos biológicos de computación.

Las mentes digitales también se beneficiarán de importantes ventajas de software:

- *Editabilidad.* Es más fácil experimentar con variaciones de parámetros en software que en

los miembros de un grupo grande, por lo menos hasta que se vuelva factible inducir la docilidad a gran escala mediante drogas o selección genética. Un "clan de copias" (un grupo de programas idénticos o casi idénticos que comparten un objetivo común) podría evitar este tipo de problemas de coordinación.

transferencia directa de memoria requiere formatos de representación normalizados. Un intercambio fácil de contenido cognitivo de alto nivel, por lo tanto, no sería posible entre cualquier par de inteligencias artificiales. En particular, no sería posible entre la primera generación de emulaciones de cerebro completo).

- *Nuevos módulos, modalidades, y algoritmos.* La percepción visual nos parece algo fácil y sin esfuerzo, muy diferente a resolver problemas de geometría, a pesar de que se necesita una cantidad masiva de computación para reconstruir, a partir de los patrones bidimensionales de estimulación de nuestras retinas, una representación tridimensional de un mundo poblado de objetos reconocibles. La razón de que esto nos parezca fácil es que hemos dedica-

ese apoyo especializado para otros ámbitos cognitivos que han llegado a ser importantes en el mundo contemporáneo —como la ingeniería, la programación de ordenadores, y la estrategia empresarial— tendría grandes ventajas sobre mentes como las nuestras que dependen de una torpe cognición de propósito general para pensar esas cosas. Nuevos algoritmos también pueden ser desarrollados para aprovechar las distintas posibilidades de hardware digital, tal como apoyo para un rápido procesamiento en serie.

Las ventajas que la inteligencia artificial, hardware y software combinados, podrían alcanzar en última instancia, son enormes.[33] Pero ¿con qué rapidez podrían alcanzarse esas ventajas potenciales? Esa es la pregunta que abordaremos ahora.

CAPÍTULO 4

La cinética de una explosión de inteligencia

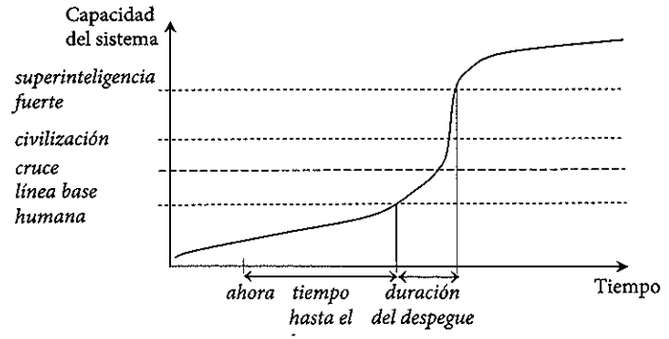


Figura 7. Configuración del despegue. Es importante distinguir entre las siguientes preguntas: "¿Se producirá un despegue? y si es así, ¿cuándo?" y "Si efectivamente se produce un despegue, ¿cómo de abrupto [illegible] qué magnitud será la fracción de la economía mundial que estará implicada en el despegue?" Estas preguntas están relacionadas pero son distintas.

Si la capacidad de razonamiento general, ¿cuánto tiempo pasará antes de que [illegible] dual y prolongada? ¿O será repentina, explosiva? [illegible] cinética de la transición a la superinteligencia como una función de la potencia de optimización y de la resistencia al progreso del sistema. Consideraremos lo que sabemos o podemos suponer razonablemente sobre el comportamiento de estos dos factores en el ámbito de la inteligencia general de nivel humano.

Sincronización y velocidad de despegue

Teniendo en cuenta que las máquinas superarán ampliamente *con el tiempo* a la biología en inteligencia general, pero que *actualmente* la cognición artificial es mucho más restringida que la cognición humana, nos preguntamos cómo de rápido se llevará a [illegible] de la cuestión que consideramos en el capítulo 1 sobre lo lejos que actualmente [illegible] la pregunta es que, *contando con que dicha máquina se desarrolle, ¿cuánto tiempo pasará a partir de entonces hasta que una máquina se vuelva radicalmente superinteligente?* Téngase en cuenta que uno puede creer que las máquinas tardarán bastante tiempo en llegar a la línea de base humana, o uno puede ser agnóstico acerca de cuánto tiempo tardarán, y, sin embargo, tener una creencia firme en que, una vez que esto suceda, el posterior ascenso hasta la superinteligencia en sentido fuerte será muy rápido.

Puede ser útil pensar esquemáticamente sobre estos asuntos, a pesar de que hacerlo implique ignorar temporalmente algunas calificaciones y detalles complicados. Consideremos, pues, un diagrama que represente gráficamente la capacidad in-

de la línea de base humana tomando como referencia cualquier métrica razonable de la capacidad intelectual general. En algún momento en el futuro, una máquina podría llegar a una paridad aproximada respecto de esta línea de base humana (que postulamos estará fijada y anclada para el año 2014, por ejemplo, aunque las capacidades de los individuos humanos deberían haber aumentado en los años intermedios): esto marcaría el inicio del despegue. Las capacidades del sistema seguirían creciendo, y en algún momento posterior, el sistema alcanzaría la paridad respecto de la capacidad combinada intelectual de toda la humanidad (de nuevo anclado en el presente): lo que podríamos llamar la "línea de base de la civilización". Eventualmente, si las capacidades del sistema siguen creciendo, alcanzaría la "superinteligencia fuerte", un nivel [illegible] de referencia que podemos llamar "el cruce", un punto más allá del cual una mejora adicional del sistema estaría impulsada principalmente por las propias acciones del sistema y no por el trabajo realizado en ella por otros.[1] (La posible existencia de un cruce como éste será importante en la subsección sobre el poder de optimización y la explosividad, más adelante en este capítulo).

Con esta imagen en mente, podemos distinguir tres clases de escenarios de transición —escenarios en los que los sistemas evolucionan de un nivel de inteligencia humana hasta una superinteligencia— en función de su brusquedad; es decir, si representan un despegue lento, rápido, o moderado.

dos poco a poco. Nuevos expertos podrían ser entrenados y acreditados. Campañas de base podrían ser movilizadas por grupos que sientan que están siendo perjudicados por el desarrollo de los acontecimientos. Si pareciera que nuevos tipos de infraestructuras más seguras o de vigilancia en masa para investigadores de IA fueran necesarias, se podrían desarrollar y desplegar tales sistemas. Naciones temerosas de una carrera armamentista de IA tendrían tiempo para tratar de negociar tratados y diseñar mecanismos que los aplicaran. La mayoría de las prepara-

horas o días. Los escenarios de despegue rápido ofrecen poca oportunidad para que los seres humanos puedan deliberar. Nadie tiene por qué siquiera notar algo inusual antes de que la partida esté ya perdida. En un escenario de despegue rápido, el destino de la humanidad depende esencialmente de los preparativos previamente establecidos. En el extremo más lento dentro del rango de escenarios de despegue rápido, algunas acciones humanas simples podrían ser posibles, como abrir de un click la "maleta nuclear"; pero dicha acción tendría que ser o bien elemental, o bien tendría que haberse planificado y preprogramado con antelación.

Moderado

Un despegue moderado es aquel que se produce en un intervalo temporal intermedio, como meses o años. Los escenarios de despegue moderados dan a los humanos alguna oportunidad de responder, pero no mucho tiempo para analizar la situación, para probar diferentes enfoques,

tentes al nuevo desafío.

patrocinado de manera encubierta por el Estado. Proyectos comerciales, pequeños equipos académicos, y grupos del tipo de "nueve hackers en un sótano" también podrían ser clandestinos; sin embargo, si la perspectiva de una explosión de inteligencia estuviera "en el radar" de los organismos de inteligencia del Estado como una prio-

de producirse, podría obstaculizar los esfuerzos para orquestar una respuesta bien coordinada; alternativamente, podría permitir soluciones más radicales de lo que circunstancias más tranquilas permitirían. Por ejemplo, en un escenario de despegue moderado donde las emulaciones baratas y capaces u otras mentes digitales fueran inundando gradualmente los mercados de trabajo en unos años, uno podría imaginar protestas masivas por parte de los trabajadores despedidos que presionan a los gobiernos para aumentar las prestaciones por desempleo o que insisten en pedir una

es lento y no moderado, pero el desequilibrio y el rápido cambio de escenarios moderados presentan oportunidades especiales para que grupos pequeños ejerzan una influencia desproporcionada.

Podría parecerle a algunos lectores que de estos tres tipos de escenario, el despegue lento es el más probable, el despegue moderado es el menos probable, y el despegue rápido es totalmente inverosímil. Podría parecer descabellado suponer que el mundo pudiera ser transformado radicalmente y que la humanidad pudiera ser derrocada de su posición como pensador alfa en el transcurso de una o dos horas. Ningún cambio de tal rapidez ha ocurrido nunca en la historia humana, y sus paralelos más cercanos —las Revoluciones Industrial y Agrícola— se desarrollaron a lo largo de escalas temporales mucho más amplias (en torno a siglos y milenios en el primer caso, en torno a décadas y siglos en el segundo). Así que la tasa base para el tipo de transición que

tema; y la capacidad de respuesta del sistema a la aplicación de una cantidad dada de tal poder de optimización. Podríamos llamar al inverso de la capacidad de respuesta "resistencia al progreso", y escribir:

particular, y la tasa de mejora que este esfuerzo produce, podríamos calcular la resistencia al progreso del sistema.

Además, se puede observar que la cantidad de energía de optimización dedicada

hacen primero, derivando en rendimientos decrecientes (aumentando la resistencia al progreso) como frutos maduros que menguan. Sin embargo, también puede haber mejoras que hagan más fácil otras mejoras adicionales, lo que conduce a cascadas de mejora. El proceso de resolver un rompecabezas comienza siendo sencillo —es fácil encontrar las esquinas y los bordes. Entonces la resistencia al progreso aumenta cuando vemos que las piezas posteriores son más difíciles de encajar. Pero a medida que el rompecabezas llega a su fin, el espacio de búsqueda va decayendo y el proceso se hace más fácil de nuevo.

Para continuar en nuestra investigación, debemos, por tanto, analizar cómo la resistencia al progreso y la optimización de potencia puede variar en los períodos críti-

Comencemos por la resistencia al progreso. Las perspectivas aquí dependen del tipo del sistema en consideración. Por completitud, primero echemos un breve vistazo a la resistencia al progreso que se encuentra a lo largo de las rutas de acceso a la superinteligencia que no implican la inteligencia artificial avanzada. Encontramos que la resistencia al progreso por esos caminos parece ser bastante alta. Una vez hecho esto, vamos a volver al caso principal, que el despegue implique inteligencia por parte de las máquinas; y allí nos encontramos con que la resistencia al progreso en el momento crítico parece baja.

Caminos que no implican la inteligencia artificial

pués volverse más problemática. El campo de la neurofarmacología todavía carece de gran parte de los conocimientos básicos necesarios para intervenir de forma competente en el cerebro sano. La desatención a la medicina mejorativa como área legítima

que se conseguirían cuando por fin el desarrollo de nootrópicos se convirtiera en algo claramente prioritario.[4]

La mejora cognitiva genética tiene un perfil de resistencia al progreso en forma de U similar a la de los nootrópicos, pero con mayores ganancias potenciales. La resistencia al progreso empezaría en un nivel alto si el único método disponible fuera la cría selectiva sostenida a lo largo de muchas generaciones, algo que obviamente es difícil de lograr a escala mundial de manera significativa. La mejora genética será más fácil a medida que la tecnología desarrollada para pruebas genéticas y selección sea barata y eficaz (particularmente cuando la selección de embriones por iteración sea

rápido que se puede progresar en el camino de la mejora genetica, sobre todo por el hecho de que las intervenciones en la línea germinal están sujetas a un retraso madurativo inevitable: esto contrarresta fuertemente la posibilidad de un despegue rápido o moderado.[5] El hecho de que la selección de embriones sólo se pueda aplicar en el contexto de la fertilización in vitro ralentizará su tasa de adopción: otro factor limitante.

La resistencia al progreso por el camino de la interfaz cerebro-ordenador parece inicialmente muy alta. En el improbable caso de que de alguna manera se convierta en algo fácil insertar implantes en el cerebro y lograr su integración funcional de alto nivel con la corteza, la resistencia al progreso podría caer en picado. A la larga, la dificultad para avanzar en este camino sería similar a la implicada en la mejora de emulaciones o

de esperar que la resistencia al progreso aumente cuando las frutas maduras (como los motores de búsqueda y el correo electrónico) vayan agotándose.

Caminos a través de la emulación y la IA

La dificultad de avanzar hacia la emulación de cerebro completo es difícil de estimar. Sin embargo, podemos señalar un hito específico del futuro: la emulación exitosa del cerebro de un insecto. Ese hito se encuentra en una colina, y su conquista dejaría a la vista gran parte del terreno por delante, lo que nos permitiría hacer una conjetura decente de la resistencia al progreso que traería consigo aumentar esta tecnología

el contrario, puede no presentar un hito o punto de observación temprano tan evidente. Es muy posible que la búsqueda de la inteligencia artificial parezca perderse en una densa selva hasta que un avance inesperado revele la línea de llegada en un claro a sólo unos pasos de distancia.

Recordemos la distinción entre estas dos preguntas: ¿Cómo de difícil es alcanzar niveles más o menos humanos de capacidad cognitiva? ¿Y cómo de difícil es llegar desde allí a niveles sobrehumanos? La primera pregunta es sobre todo relevante para predecir cuánto tiempo pasará antes del inicio del despegue. La segunda pregunta es la clave para evaluar la forma del despegue, que es nuestro objetivo ahora. Y aunque podría ser tentador suponer que el paso desde el nivel humano al nivel sobrehumano

Consideremos en primer lugar la emulación de cerebro completo. Las dificultades para la creación de la primera emulación humana son de un tipo muy diferente de aquellas implicadas en la mejora de una emulación existente. La creación de una primera emulación implica enormes desafíos tecnológicos, particularmente en lo que se refiere al desarrollo de las necesarias capacidades de escaneo y de interpretación de imágenes. Este paso también podría requerir cantidades considerables de capital físico —un parque mecánico de escala industrial con cientos de máquinas de escaneo de alto rendimiento no es inverosímil. Por contraste, la mejora de calidad de una emulación existente implica afinar algoritmos y estructuras de datos: es esencialmente un problema de software, y podría resultar ser mucho más fácil que el perfeccio-

mucha atención durante la fase de desarrollo a la eficiencia de la puesta en práctica

y podrían estar disponibles oportunidades sencillas para el ahorro de eficiencia computacional. (Una reorganización arquitectónica más fundamental también podría ser posible, pero eso nos llevaría fuera del camino de la emulación y hacia el territorio de la IA).

Otra forma de mejorar el código base una vez que la primera emulación se hubiera producido es escanear cerebros adicionales con diferentes o superiores habilidades y talentos. También se produciría un crecimiento en la productividad como consecuencia de la adaptación a estructuras organizativas y flujos de trabajo de los atributos específicos de las mentes digitales. Puesto que no hay precedentes en la economía humana de un trabajador que pueda literalmente ser copiado, restablecido y puesto

a funcionar a diferentes velocidades, los coordinadores de la primera población de

zado, las variaciones algorítmicas más prometedoras se habrán probado y las oportunidades más fáciles para la innovación organizacional se habrán aprovechado. La biblioteca de plantillas se habrá ampliado de forma que la adquisición de más escáneres cerebrales añadirá poco beneficio al trabajo con las plantillas existentes. Ya que una plantilla se puede multiplicar, cada copia puede ser entrenada individualmente en un campo diferente, y esto se puede hacer a una velocidad electrónica, por lo que podría ser que el número de cerebros que tuvieran que ser escaneados con el fin de capturar la mayor parte del potencial de ganancia económico fuera pequeño. Posiblemente un solo cerebro sería suficiente.

Otra posible causa de la escalada de resistencia al progreso es la posibilidad de que

posible, sin embargo, que los acontecimientos políticos vayan en la dirección opuesta, lo que contribuiría a una caída en la resistencia al progreso. Esto puede suceder si una restricción inicial en el uso de mano de obra de emulación da paso a una explotación sin trabas, ya que la competencia aumentaría y los costes económicos y estratégicos del terreno moral se habrían despejado.

En cuanto a la inteligencia artificial (una máquina inteligente, no una emulación), la dificultad de elevar un sistema desde el nivel humano hasta la inteligencia sobrehumana mediante mejoras algorítmicas depende de los atributos del sistema en particular. Diferentes arquitecturas pueden tener una resistencia al progreso muy diferente.

Potencia de optimización y explosividad

Habiendo examinado la cuestión de la resistencia al progreso ahora debemos recurrir a la otra mitad de la ecuación esquemática, la *potencia de optimización*. Para refrescar la memoria: *Tasa de cambio de Inteligencia = potencia de optimización / resistencia al progreso*. Como se refleja en este esquema, un despegue rápido no requiere que la resistencia al progreso en la fase de transición sea baja. Un despegue rápido también podría suceder si la resistencia al progreso es constante o incluso moderadamente creciente, siempre que la potencia de optimización enfocada a mejorar el rendimiento

nas razones para pensar que la potencia optimización aplicada *aumentará* durante la transición, al menos en ausencia de medidas deliberadas para evitar que esto suceda.

Podemos distinguir dos fases. La primera fase comienza con el inicio del despegue, cuando el sistema alcanza la línea base de la inteligencia humana individual. Como la capacidad del sistema sigue en aumento, podría usar parte o la totalidad de esa capacidad para mejorarse a sí misma (o para diseñar un sistema sucesor que, a nuestros efectos, viene a ser lo mismo). Sin embargo, la mayor parte de la potencia de optimización aplicada al sistema aún provendría de fuera del sistema, ya sea del trabajo de los programadores e ingenieros que trabajan en el proyecto o del tipo de trabajo realizado por el resto del mundo que pueda ser incorporado y utilizado por el proyecto.[16]

nes tanto desde dentro del proyecto como desde el mundo exterior son propensas a aumentar a medida que las probabilidades de éxito del enfoque elegido se hacen manifiestas. Los investigadores trabajarían más duro, se reclutarían más investigadores y se compraría más potencia de cálculo para acelerar el progreso. El aumento podría ser especialmente dramático si el desarrollo de la inteligencia artificial de nivel humano tomara al mundo por sorpresa, en cuyo caso lo que antes hubiera sido un pequeño proyecto de investigación, pronto podría convertirse en el foco de intensos esfuerzos de investigación y desarrollo por parte de todo el mundo (aunque algunos de esos esfuerzos podrían ser canalizados hacia proyectos de la competencia).

Una segunda fase de crecimiento se iniciará si en algún momento el sistema adquiere tanta capacidad como para que la mayor parte de la potencia de optimización ejercida en él venga del propio sistema (marcado por el nivel variable denominado

porcional de la cantidad de energía que se aplica a la optimización de sus ulteriores mejoras. Si la resistencia al progreso permaneciera constante, esta retroalimentación dinámica produciría un crecimiento exponencial (véase el cuadro 4). Esta constante duplicadora depende del escenario, pero podría ser muy corta —de unos pocos segundos en algunos escenarios— si el crecimiento se produjera a velocidades electrónicas, que pueden acontecer como resultado de mejoras algorítmicas o de la explotación de un excedente de contenidos o de hardware.[17] El crecimiento impulsado por construcción física, tal como la producción de nuevos ordenadores o equipo de

Cuadro 4. Sobre la cinética de una explosión de inteligencia

Podemos escribir la tasa de cambio en la inteligencia como la relación entre la potencia aplicada a la optimización del sistema y la resistencia al progreso del sistema:

$$\frac{dI}{dt} = \frac{\mathfrak{D}}{\mathfrak{R}}$$

La cantidad de energía que actúa sobre la optimización de un sistema es la suma de la

zación desde fuera. Por ejemplo, una IA seminal podría ser mejorada a través de una combinación de sus propios esfuerzos y los esfuerzos de un equipo de programación humano, y quizás también a través de los esfuerzos de la comunidad global más amplia de investigadores que hacen continuos avances en la industria de semiconductores, de ciencias de la computación, y de campos relacionados[19]

$$\mathfrak{D} = \mathfrak{D}_{sistema} + \mathfrak{D}_{proyecto} + \mathfrak{D}_{mundo}$$

Una IA seminal comienza con capacidades cognitivas muy limitadas. Al principio, por lo tanto, el $\mathfrak{D}_{sistema}$ es pequeño.[20] ¿Qué ocurre con $\mathfrak{D}_{proyecto}$ y $\mathfrak{D}_{mundo}$? Hay casos en los que un proyecto tiene una capacidad más grande que el resto del mundo combinado —el

proyecto contenga solo una pequeña parte de la capacidad de investigación relevante de todo el mundo. Pero incluso en casos en los que el mundo exterior tenga una mayor cantidad total de capacidad de investigación relevante que cualquier proyecto, $\mathfrak{D}_{proyecto}$, no obstante, podría superar a $\mathfrak{D}_{mundo}$, ya que gran parte de la capacidad del mundo exterior no se centrará en el sistema en cuestión. Si un proyecto comenzara a parecer prometedor —lo cual ocurriría cuando un sistema pase la línea base humana, si no antes— éste podría atraer más inversión, aumentando el $\mathfrak{D}_{proyecto}$. Si los logros del proyecto son públicos, $\mathfrak{D}_{mundo}$ también podría elevarse a medida que los avances inspiraran un mayor interés en la inteligencia artificial en general y diversos poderes intentaran entrar en el juego. Durante la fase de transición, por lo tanto, la potencia de optimización total aplicada a la mejora de un sistema cognitivo es probable que aumente la capacidad de realizar aumentos en el sistema.[21]

potencia de optimización que se le aplica desde el exterior (en todas las dimensiones importantes de mejora):

$$\mathfrak{D}_{sistema} > \mathfrak{D}_{proyecto} > \mathfrak{D}_{mundo}$$

Este *cruce* es significativo porque más allá de este punto, una mejora adicional de las capacidades del sistema contribuye fuertemente a aumentar la potencia de optimización total aplicada a la mejora del sistema. Con ello entramos en un régimen continuo de fuerte auto-mejora recursiva. Esto conduciría a un crecimiento explosivo de la capacidad del

Cuadro 4. *Continúa*

Como ejemplo, consideremos primero un escenario en el que la resistencia al progreso fuera constante, de modo que la tasa de aumento en inteligencia de una IA fuera igual a la potencia de optimización que se le aplicara. Supongamos que toda la potencia

gencia a la tarea de amplificar su propia inteligencia, para que $\approx_{sistema} = 1$. A continuación, tendríamos

$$\frac{dI}{dt} = \frac{I}{k}$$

Resolviendo esta simple ecuación diferencial se obtiene la función exponencial

$$I = Ae^{t/k}$$

Pero que la resistencia al progreso sea constante es un caso bastante especial. La resistencia al progreso probablemente podría decaer alrededor de la línea de base humana, debido a uno o más de los factores mencionados en el apartado anterior, y permanecer por debajo del cruce y a una cierta distancia más allá de él (tal vez hasta que el sistema se acerque finalmente a límites físicos fundamentales). Por ejemplo, supongamos que la potencia de optimización aplicada al sistema es aproximadamente constante (es decir

ma cada 18 meses. (Esto estaría más o menos en línea con el ritmo de mejoras históricas de la ley de Moore, combinada con avances en software23). Esta tasa de mejora, si se logra a través de una potencia de optimización más o menos constante, implicaría que la resistencia al progreso decaería de manera inversa a la potencia del sistema:

$$\frac{dI}{dt} = \frac{(c + I)}{1/I} = (c + I)I.$$

Si la resistencia al progreso siguiera cayendo a lo largo de este patrón hiperbólico, entonces, cuando la IA alcanzara el punto de cruce, la cantidad total de energía de optimización aplicada a la mejora de la IA se habría duplicado. Tendríamos entonces

$$dI/dt = (c + I)/(1/I) = (c + I)\,I$$

La siguiente duplicación se produciría 7,5 meses más tarde. En 17,9 meses, la capacidad del sistema se habría multiplicado por mil, obteniendo así una superinteligencia de velo-

Esta particular trayectoria de crecimiento nos llevaría a una singularidad positiva en t = 18 meses. En realidad, el supuesto de que la resistencia al progreso sería constante

procesamiento de información, si no antes.

Cuadro 4. *Continúa*

del punto de cruce contribuye con fuerza a hacer que el despegue sea más rápido de lo que habría sido de otro modo.

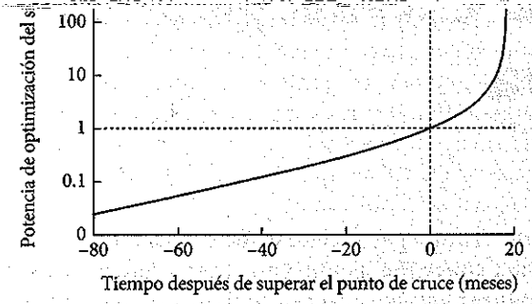


Figura 9. Un sencillo modelo de una explosión de inteligencia.

Por lo tanto, es probable que la potencia de optimización aplicada aumentara du-

tarde, porque la propia inteligencia artificial se volvería capaz de impulsar un mayor progreso a velocidades digitales. Esto crearía una posibilidad real de despegue rápido o medio, *incluso si la resistencia al progreso fuera constante o ligeramente creciente en toda la línea base humana.*[18] No obstante, vimos en el apartado anterior que hay factores que podrían conducir a una gran caída en la resistencia al progreso en torno a la capacidad de línea base humana. Estos factores incluyen, por ejemplo, la posibilidad de expansión rápida de hardware una vez que se haya alcanzado una mente-software de trabajo; la posibilidad de mejoras algorítmicas; la posibilidad de escanear cerebros adicionales (en el caso de la emulación de cerebro completo); y la posibilidad de incorporar rápidamente grandes cantidades de contenido al digerir internet (en el caso de la inteligencia artificial).[24]

A pesar de estas observaciones, la forma de la curva de resistencia al progreso en la región clave aún no está bien caracterizada. En particular, no está claro lo difícil que sería mejorar la calidad del software de emulación de nivel humano o de la IA.

tar la potencia computacional de un proyecto pequeño mediante el gasto de mucho veces más dinero en potencia de cálculo, o esperando unos años para que el precio de

las computadoras cayera, es posible que la primera inteligencia artificial que llegue a

CAPÍTULO 5

Ventaja estratégica decisiva

Una pregunta distinta pero relacionada con la cuestión de la cinética es ¿habrá un poder superinteligente o muchos? ¿Podría una explosión de inteligencia impulsar un proyecto tan adelantado respecto del resto como para que fuera capaz de dictar el futuro? ¿O será el progreso más uniforme, desplegándose en un frente amplio, con muchos proyectos participando pero sin que ninguno se asegure una ventaja abrumadora y permanente?

En el capítulo anterior se analizó un parámetro clave para determinar el tamaño de la brecha que podría abrirse entre un poder puntero y sus competidores más cercanos —a saber, la velocidad de transición desde una inteligencia humana a una inteligencia

pegue fuera *rabido* (completado en el transcurso de horas, días o semanas), entonces

mo tiempo: casi con certeza, el primer proyecto habría completado su despegue antes de que ningún otro proyecto hubiera comenzado su propio despegue. Si el despegue fuera *lento* (prolongándose durante muchos años o décadas), entonces podría haber múltiples proyectos despegando simultáneamente, de modo que aunque los proyectos hacia el final de la transición hubieran ganado enormemente en capacidad, no habría ningún momento en el que algún proyecto llegara a estar lo suficientemente lejos delante de los demás como para tener una ventaja abrumadora. Un despegue de velocidad *moderada* se posicionaría entre los anteriores, con una de estas posibilidades como condición: que pudiera haber o no más de un proyecto despegando al mismo tiempo.[1]

¿Llegaría un proyecto de inteligencia artificial a adelantarse tanto a la competencia como para conseguir una *ventaja estratégica decisiva* —es decir, un nivel de ventajas tecnológicas y otro tipo de ventajas suficientes como para alcanzar la dominación

utilizada para eliminar a los competidores y formar una *Unidad* (un orden mundial

¿Conseguirá el proyecto adelantado una ventaja estratégica decisiva?

Un factor que influye en el tamaño de la brecha entre el proyecto adelantado y sus perseguidores es la velocidad de difusión de aquello que da al líder una ventaja competitiva. Un proyecto adelantado podría tener dificultades para obtener y mantener una ventaja grande si los perseguidores pudieran copiar fácilmente sus ideas e innovaciones. La imitación crea un viento de frente que perjudica a los líderes y beneficia a los rezagados, especialmente si la propiedad intelectual está débilmente protegida. Un proyecto adelantado también puede ser especialmente vulnerable a la expropiación

namente a medida que crezca la brecha entre el proyecto adelantado y sus perseguidores. Así como un ciclista que se descuelga a demasiada distancia del pelotón ya no está protegido del viento por los ciclistas de delante, de igual modo un perseguidor tecnológico que se quedara bastante rezagado respecto de la primera línea podría encontrar dificultades para asimilar los avances que se realizaran en los primeros puestos.[2] La brecha en comprensión y capacidad podría haber crecido demasiado. El líder podría haber migrado a una plataforma tecnológica más avanzada, por lo que las innovaciones posteriores serían intransferibles a las plataformas primitivas utilizadas por los rezagados. Un líder lo suficientemente destacado podría tener la capacidad de frenar la fuga de información de sus programas de investigación y de sus instalaciones más delicadas, o la capacidad de sabotear los esfuerzos de sus competidores por desarrollar sus propias capacidades avanzadas.

'estan contrarrestadas por las ineficiencias burocráticas y problemas administrativos, incluyendo dificultades como la de mantener el secreto comercial.[3] Estos problemas presumiblemente limitarían el crecimiento de un proyecto de inteligencia artificial siempre que fuera manejado por seres humanos. Un sistema de inteligencia artificial, sin embargo, podría evitar algunas de estas ineficiencias de escala, ya que los módulos de la IA (en contraste con los de los trabajadores humanos) no tienen por qué tener preferencias individuales que difieran de los del sistema en su conjunto. Así, el sistema de inteligencia artificial podría evitar una parte considerable de las ineficiencias derivadas de los problemas administrativos de las empresas humanas. La misma ventaja —tener partes completamente leales— también haría que fuera más fácil para un sistema de inteligencia artificial conseguir metas clandestinas de largo alcance. Una IA no tendría empleados descontentos predispuestos a ser captados por los competidores o a ser sobornados para convertirse en informadores.[4]

habitual para proyectos tecnológicos estratégicamente importantes está entre unos

Cuadro 5. Carreras tecnológicas: algunos ejemplos históricos

A lo largo de grandes períodos de tiempo histórico, se ha producido un aumento en la [illegible]

[illegible]

guardado. Revelar las técnicas era castigable con la muerte, al igual que exportar los gusanos de seda o sus huevos fuera de China. Los romanos, a pesar del alto precio que conllevaba importar paños de seda en su imperio, nunca aprendieron el arte de fabricación de la seda. No fue hasta alrededor del año 300 cuando una expedición japonesa logró capturar algunos huevos de gusanos de seda junto con cuatro niñas chinas jóvenes, que se vieron obligadas a divulgar el arte a su secuestradores.[6] Bizancio se unió al club de los productores en el año 522. La historia de la porcelana también cuenta con grandes retrasos temporales. El arte fue practicado en China durante la dinastía Tang alrededor del año 600 d.C. (y podría haber estado en uso ya desde el 200 d.C.), pero fue dominado por los europeos sólo en el siglo XVIII.[7] Los vehículos de ruedas aparecieron en varios sitios de toda Europa y Mesopotamia alrededor del 3500 a.C., pero llegó a las Américas [illegible]

[illegible]

Europa y América del Norte en sólo un año!).

La competencia tecnológica se ha estudiado extensamente, particularmente en el contexto de las carreras de patentes y las carreras armamentísticas.[9] Está más allá del alcance de nuestra investigación revisar esa literatura aquí. Sin embargo, es instructivo examinar algunos ejemplos de carreras tecnológicas estratégicamente importantes del siglo XX (véase el cuadro 7).

Con respecto a estas seis tecnologías que las superpotencias rivales consideraron de importancia estratégica por su relevancia militar o simbólica, las brechas entre el líder y el rezagado más cercano fueron (de manera muy aproximada) 49 meses, 36 meses, 4 meses, 1 mes, 4 meses, y 60 meses, respectivamente —superiores a la duración de un

Cuadro 5. *Continúa*

…ativas

India	Israel	Pakistán
1974	1979?	1998
1998	?	—
1980	1988	—
—	—	—
2012	2008	—[16]
2014[19]	2008?	

…ras tecnológicas estratégica…

Estados Unidos	Unión Soviética	Reino Unido
1945	1949	1952
1952	1953[11]	1957
1958	1957	1971
1961	1961	—
1959	1960	1968[15]
1970	1975	1979

Cuadro 5. *Continúa*

disponible; excepto en algunos casos, cuando parece que el descubrimiento ha de ofrecer una ventaja estratégica, y la publicación se retrasa. Por ejemplo, dos de las ideas más importantes de la criptografía de clave pública son el protocolo de intercambio de claves Diffie-Hellman y el esquema de cifrado RSA. Estos fueron descubiertos por la comu-

podrían ofrecer una analogía más estrecha con los proyectos de IA, pero es más difícil dar ejemplos nítidos de retrasos típicos ya que el software se suelen dar a conocer de manera incremental y las funcionalidades de los sistemas de la competencia a menudo no son directamente comparables.

Incluso en ausencia de dinámicas que condujeran a un efecto de bola de nieve, algunos proyectos acabarían naturalmente con un mejor personal de investigación, con mayor liderazgo, y con mejor infraestructura, o simplemente se tropezarían con mejores ideas. Si dos proyectos persiguen enfoques alternativos, uno de los cuales resulta funcionar mejor, es posible que los proyectos rivales tarden muchos meses en cambiar al enfoque superior incluso aunque fueran capaces de seguir de cerca lo que el líder estuviera haciendo.

un despegue lento, es muy probable que varios proyectos llevaran a cabo el proceso en paralelo. Pero el análisis necesita ir un paso más allá. La pregunta clave no es el número de proyectos que despegarían en tándem, sino cuántos proyectos emergerían lo suficientemente bien capacitados para que ninguno de ellos tuviera una ventaja estratégica decisiva. Si el proceso de despegue fuera relativamente lento al principio y luego se acelerara, la distancia entre los proyectos que compiten tendería a crecer. Regresando a nuestra metáfora de la bicicleta, la situación sería análoga a la de un par de ciclistas pedaleando en una colina empinada, uno detrás del otro a cierta distancia —la brecha entre ellos crecería cuando el ciclista adelantado superara la cima y co-

daría al proyecto líder una ventaja estratégica decisiva y la oportunidad de valerse de su liderazgo para obtener el control permanente mediante la desactivación de los proyectos en competencia y el establecimiento de una Unidad. (Nótese que el concepto de Unidad es uno abstracto: una Unidad podría ser una democracia, una tiranía, una sola IA dominante, un sólido conjunto de normas globales que incluyeran disposiciones eficaces para su cumplimiento, o incluso un gobernante supremo alienígena —su definición característica es simplemente que es algún tipo de administración que

sitiva de la potencia optimizadora entre en juego, un escenario de este tipo es una posibilidad seria, y aumenta las posibilidades de que el proyecto que lleve la delantera alcance una ventaja estratégica decisiva, aunque el despegue no fuera rápido.

¿Qué magnitud tendrá el proyecto exitoso?

Algunas rutas de acceso a la superinteligencia requieren de grandes recursos y, por lo tanto, es probable que sean exclusivamente proyectos grandes y bien financiados. La emulación de cerebro completo, por ejemplo, requiere diferentes tipos de conocimientos y un montón de equipo. Las mejoras en inteligencia biológica y en interfaces cerebro-ordenador también tendrían un factor de gran escala: mientras que una pequeña empresa de biotecnología podría inventar una o dos drogas, logrando la super-

inteligencia colectiva haciendo más eficientes las organizaciones y redes requiere una financiación aún más extensa, que incluiría la participación de gran parte de la economía mundial.

El camino de la IA es más difícil de evaluar. Tal vez sería necesario un programa de investigación muy grande; tal vez podría ser realizado por un grupo pequeño. Un escenario de pirata informático solitario tampoco puede excluirse. La construcción de una IA seminal podría requerir conocimientos y algoritmos desarrollados a lo largo de muchas décadas por la comunidad científica mundial. Pero es posible que la última idea crítica en el avance pudiera provenir de una sola persona o de un pequeño

Hay que distinguir la pregunta sobre lo grande que será el proyecto que directamente *construya* el sistema, de la pregunta sobre lo grande que será el grupo que *controle* si, cómo y cuándo se crea el sistema. La bomba atómica fue creada principalmente por un grupo de [illegible] a alrededor de 130.000 personas en su apogeo, la gran mayoría de los cuales eran [illegible][23]. Estos expertos técnicos, sin embargo, fueron controlados por el ejército estadounidense, que a su vez estaba dirigido por el gobierno de los Estados Unidos, que en última instancia era responsable ante el electorado estadounidense, que en su momento constituía aproximadamente una décima parte de la población adulta mundial.[24]

Monitorización

Dadas las implicaciones de seguridad extrema de la superinteligencia, los gobiernos probablemente traten de nacionalizar cualquier proyecto de su territorio que ellos consideren cercano a lograr un despegue. Un Estado poderoso también podría intentar adquirir proyectos ubicados en otros países a través del espionaje, el robo, el [illegible] la conquista militar o cualquier otro medio disponible. Un Estado de gran alcance que no pudiera adquirir un proyecto extranjero podría, en su lugar, destruirlo, sobre todo si el país receptor careciera de un medio eficaz de disuasión. Si las estructuras de gobierno [illegible] momento en que un avance comenzara a parecer inminente, es posible que los proyectos prometedores fueran puestos bajo control internacional.

Una cuestión importante, por lo tanto, es si las autoridades nacionales o internacionales serían capaces de ver venir una explosión de inteligencia inminente. En la actualidad, las agencias de inteligencia no parecen estar buscando con mucho ahínco proyectos de IA prometedores u otras formas de amplificación de la inteligencia.[25] Si de hecho no prestan (mucha) atención, esto se debe probablemente a la percepción ampliamente compartida de que no hay ninguna posibilidad de que surja una superinteligencia de manera inminente. Si llegara a convertirse en una creencia común entre los científicos prestigiosos el que hay una posibilidad sustancial de que la superinteligencia estuviera a la vuelta de la esquina, las principales agencias de inteligencia del mundo probablemente comenzarían a monitorizar grupos e individuos [illegible] comenzara a mostrar un progreso suficiente podría entonces ser nacionalizado rápidamente. Si las élites políticas se persuadieran de la gravedad del riesgo, los esfuerzos civiles en áreas delicadas podrían ser regulados o prohibidos.

¿Cómo de difícil será esa monitorización? La tarea es más fácil si el objetivo es sólo no perder de vista al proyecto puntero. En ese caso, una vigilancia centrada en los proyectos mejor dotados de recursos puede ser suficiente. Si el objetivo es, en cambio, evitar que cualquier trabajo tenga lugar (al menos fuera de las instituciones especialmente autorizadas), entonces la vigilancia tendría que ser más amplia, ya que muchos proyectos pequeños e individuos estarán en condiciones de hacer al menos algunos progresos [illegible] dor personal y, por lo tanto, sería más difícil de controlar. Parte del trabajo teórico se podría hacer con papel y lápiz. Aun así, no sería demasiado difícil de identificar a las personas más capaces, con un serio y documentado interés en investigaciones sobre [illegible] inteligencia artificial general. Estas personas suelen dejar rastros visibles. Es posible que hayan publicado artículos académicos, presentado en congresos, publicado en [illegible] informática. También pueden haber tenido comunicaciones con otros investigadores de IA, lo que nos permitiría identificarlos mediante mapeados de gráfica social.

Los proyectos diseñados desde un principio para ser secretos podrían ser más difíciles de detectar. Un proyecto común de desarrollo de software podría servir como avanzadilla.[26] Sólo un cuidadoso análisis del código que estuvieran produciendo revelaría la verdadera naturaleza de lo que el proyecto estaría tratando de lograr. Tal análisis requeriría una gran cantidad de mano de obra (altamente cualificada), por lo que sólo un pequeño número de proyectos sospechosos de este tipo podrían ser analizados. La tarea sería mucho más fácil si una tecnología de detección de mentiras efectiva se hubiera desarrollado y pudiera ser utilizada rutinariamente en este tipo de vigilancia.[27]

Otra razón por la que los Estados podrían dejar escapar acontecimientos precur- [illegible] sores es la dificultad inherente de la previsión de algunos tipos de avance. Esto es más relevante para la investigación en IA que para el desarrollo de la emulación de cerebro [illegible] catenación de avances clave.

También es posible que las agencias de inteligencia y otras administraciones gubernamentales sean algo torpes o rígidas y que esto les impida comprender el significado de algunos acontecimientos que podrían ser evidentes para algunos grupos externos. Las dificultades para la comprensión oficial de una potencial explosión de inteligencia podrían ser especialmente graves. Es concebible, por ejemplo, que el tema se inflame con controversias religiosas o políticas, convirtiéndolo en un tabú para los funcionarios de algunos países. El tema podría llegar a ser asociado con alguna figura desacreditada o con la charlatanería y la publicidad en general, y, por lo tanto, rechazado por los científicos respetados y otras figuras de la clase dirigente. (Como vimos en el capítulo 1, algo similar ya ha sucedido dos veces: recordemos los dos "inviernos de IA".) Los grupos industriales podrían ejercer presión política para evitar calumnias para marginar a los que expresaran sus preocupaciones acerca de las consecuencias a largo plazo de la ciencia que se estuviera haciendo.[28]

En consecuencia, un fracaso total de la inteligencia no puede descartarse. Tal falta es especialmente probable si los avances se produjeran en el futuro más próximo, antes de que el tema se haya elevado al debate público. E incluso si las agencias de inteligencia no se equivocan, los líderes políticos podrían no escuchar o actuar según sus consejos. Dar comienzo al Proyecto Manhattan requirió un esfuerzo extraordinario por parte de varios físicos visionarios, incluyendo especialmente a Mark Oliphant y Leó Szilárd: este último convenció a Eugene Wigner de que convenciera a Albert Einstein para poner su nombre en una carta dirigida a convencer al presidente

Para bien o para mal, probablemente sería más difícil que un pequeño grupo de activistas influyeran en una explosión de inteligencia si grandes jugadores, como los Estados, estuvieran tomando parte activa. Las oportunidades para que individuos particulares reduzcan la cantidad global de riesgo existencial de una potencial explosión de inteligencia son, por tanto, mayores en escenarios en los que los grandes jugadores siguen siendo relativamente ajenos a la cuestión, o en los que los primeros esfuerzos de los activistas marquen una gran diferencia sobre la posibilidad, el cuándo, los medios, o la actitud con que los grandes jugadores entren en el juego. Por lo tanto, los activistas que busquen el máximo impacto tal vez deseen centrar la mayor parte de su

en los que los grandes jugadores terminan acertando son más probables.

Colaboración internacional

La coordinación internacional es más probable si las estructuras de gobierno global se hacen más fuertes. La coordinación también podría ser más probable si el significado de una explosión de inteligencia es muy apreciado antes de tiempo y si la supervisión efectiva de todos los proyectos serios es factible. Incluso si la monitorización no es factible, no obstante, la cooperación internacional seguiría siendo posible. Muchos países podrían unirse para apoyar un proyecto conjunto. Si tal proyecto conjunto tuviera recursos suficientemente buenos, podría tener una buena oportunidad de ser el primero en llegar a la meta, sobre todo si cualquier proyecto rival se ve obligado a ser pequeño y secreto para eludir la detección.

Existen precedentes de exitosas colaboraciones científicas multinacionales a gran

tados Unidos fue un objetivo importante en sí mismo[30]). El logro de una colaboración similar en un proyecto con enormes implicaciones en seguridad sería más difícil. Un país que se creyera capaz de lograr un avance unilateral podría tener la tentación de ir por su cuenta en vez de subordinar sus esfuerzos a un proyecto conjunto. Un país también podría abstenerse de unirse a una colaboración internacional por temor a que otros participantes pudieran desviar ideas generadas en colaboración y utilizarlas para acelerar un proyecto nacional encubierto.

Un proyecto internacional necesitaría, por lo tanto, sobreponerse a los principales problemas de seguridad, y probablemente se necesitaría una buena cantidad de con-

tales y defenderse frente a las naciones más pequeñas que pudieran desarrollar armas nucleares. Sin embargo, Gorbachov no fue persuadido por esta aparente propuesta inmejorable. Consideraba la táctica como un ardid, negándose a creer que los estadounidenses fueran a compartir los frutos de su investigación militar más avanzada en un momento en que ni siquiera estaban dispuestos a compartir con los soviéticos su tecnología para ordeñar vacas.[31] Independientemente de si Reagan fue, de hecho, sincero en su oferta de colaboración entre superpotencias, la desconfianza hizo que la propuesta no pudiera ver la luz.

La colaboración es más fácil de lograr entre aliados, pero incluso allí no es auto-

durante la Segunda Guerra Mundial, Estados Unidos ocultó su proyecto de bomba atómica a la Unión Soviética. Los Estados Unidos sí colaboraron en el Proyecto Man-

Unión Soviética su éxito al descifrar el código alemán Enigma, pero lo compartió —aunque no sin reticencias— con Estados Unidos.[33] Esto sugiere que a fin de lograr colaboración internacional sobre alguna tecnología de importancia fundamental para la seguridad nacional, podría ser necesario haber construido previamente una relación cercana y de confianza.

Volveremos en el capítulo 14 sobre la conveniencia y viabilidad de la colaboración internacional en el desarrollo de tecnologías de amplificación de la inteligencia.

De una ventaja estratégica decisiva a la Unidad

¿Decidiría un proyecto que obtuviera una ventaja estratégica decisiva usarla para for-

—y por algún tiempo más— los Estados Unidos pudieron haber tenido, o estuvieron en condiciones de lograr, una ventaja militar decisiva.

Los Estados Unidos, entonces, teóricamente podrían haber utilizado su monopolio nuclear para crear una Unidad. Una forma en la que podrían haberlo hecho habría sido embarcándose en un enorme esfuerzo por construir un gran arsenal nuclear y amenazando (y si es necesario, llevando a cabo) un primer ataque nuclear para destruir la capacidad industrial de cualquier programa nuclear incipiente en la URSS y en cualquier otro país tentado a desarrollar capacidad nuclear.

Un curso más benigno de acción, que también podría haber tenido la oportunidad

negociar un gobierno internacional fuerte —unas Naciones Unidas sin derecho a veto

de juegos y uno de los arquitectos de la estrategia nuclear estadounidense).[34] Tal vez es un signo de progreso civilizatorio que la misma idea de amenazar con un ataque nuclear preventivo nos parezca hoy algo estúpido o moralmente obsceno.

Una versión más benigna del enfoque fue ensayada en 1946 por los Estados Unidos, [illegible] tecnología nuclear serían puestos bajo el control de un organismo internacional que funcionaría bajo los auspicios de las Naciones Unidas. La propuesta pedía que los miembros permanentes del Consejo de Seguridad renunciaran a sus vetos en asuntos relacionados con las armas nucleares con el fin de prevenir que cualquier gran poder que no respetara el acuerdo vetara la imposición de sanciones.[35] Stalin, al ver que la Unión Soviética y sus aliados podrían perder fácilmente la votación en el Consejo de Seguridad y en la Asamblea General, rechazó la propuesta. Una atmósfera helada de sospecha mutua descendió sobre las relaciones entre los antiguos aliados de guerra, una desconfianza que luego se solidificó en la Guerra Fría. Como había sido ampliamente predicho, a estos acontecimientos le siguieron una carrera armamentista nuclear costosa y extremadamente peligrosa.

Hay muchos factores que pueden disuadir a una organización humana en poder de una ventaja estratégica decisiva de la creación de una Unidad. Estos factores incluyen funciones no agregativas o delimitadas de servicios públicos, reglas de decisión no-maximizantes, la confusión y la incertidumbre, problemas de coordinación, y diversos costes asociados a la toma de posesión. Pero ¿y si no fuera una organización humana, sino un agente artificial superinteligente el que llegara a tener en posesión una ventaja estratégica decisiva? ¿Los factores antes mencionados serían igualmente eficaces en impedir que una IA intentara tomar el poder? Recorramos brevemente la lista de factores y consideremos cómo podrían aplicarse en este caso.

Los individuos humanos y las organizaciones humanas suelen tener preferencias sobre los recursos que no están bien representados por una "función de utilidad sin [illegible] ciento de expandirse por diez. Para los individuos y los gobiernos, hay rendimientos decrecientes en la mayoría de los recursos. La misma necesidad podría *no* aplicarse para las IAs. (Volveremos al problema de la motivación de la IAs en los capítulos siguientes). Por lo tanto, una IA podría ser más propensa a seguir un curso de acción arriesgado que tuviera alguna posibilidad de darle el control del mundo.

Los seres humanos y las organizaciones humanas también podrían operar con procesos de toma de decisiones que no buscaran maximizar la utilidad esperada. Por ejemplo, podrían basarse en reglas de decisión guiadas por la aversión al riesgo fundamental, o por "ciega satisfacción" que se centran en cumplir con ciertos umbrales de adecuación o con limitaciones "deontológicas" secundarias que proscriben ciertos

estratégica y la incertidumbre. Las revoluciones, incluso cuando tienen éxito en el derrocamiento del orden existente, a menudo no producen el resultado que sus instigadores habían prometido. Esto tiende a detener la mano de un agente humano si la acción contemplada es irreversible, contraria a la legalidad, y carece de precedentes [illegible] se propusiera utilizar su aparente ventaja estratégica decisiva para consolidar su posición dominante.

Otro factor importante que puede impedir a grupos la explotación de una potencial ventaja estratégica decisiva es el problema de la coordinación interna. Los miembros de una conspiración que estuviera en condiciones de tomar el poder deben preocuparse no sólo por tener infiltrados del exterior, sino también de ser derrocado por alguna coalición interna en poder de información privilegiada. Si un grupo está formado por un centenar de personas, y una mayoría de sesenta puede tomar el poder y privar de derechos a los no-conspiradores, ¿qué implicaría entonces detener un subconjunto de treinta y cinco de estos sesenta e impedirles que priven de sus derechos de voto a los [illegible] derechos de voto a otros quince? Cada uno de los cien originales podría tener buenas razones para mantener ciertas normas establecidas que previnieran la desintegración general a la que podría llevar cualquier intento de cambiar el contrato social por medio de una toma de poder directo. Este problema de coordinación interna no se aplica a un sistema de inteligencia artificial que estuviera constituida por un sólo agente unificado.[36]

Por último, está la cuestión del coste. Incluso si Estados Unidos pudiera haber utilizado su monopolio nuclear para establecer una Unidad, podría no haber sido capaz de hacerlo sin incurrir en costes sustanciales. En el caso de un acuerdo negociado para poner las armas nucleares bajo el control de unas Naciones Unidas reformadas y reforzadas, estos costes podrían haber sido relativamente pequeños; pero los costes —morales, económicos, políticos y humanos— de intentar efectivamente la conquista [illegible] [illegible], por ejemplo, un escenario en el que una nación tuviera una gran ventaja tecnológica que fuera capaz de desarmar de forma segura a todas las demás naciones con sólo pulsar un botón, sin que nadie muriera o fuera herido, y casi sin daños infraestructurales o al medio ambiente. Con tal superioridad tecnológica casi mágica, un ataque preventivo sería mucho más tentador. O pensemos en un nivel aún mayor de superioridad tecnológica que pudiera llevar a otras naciones a dejar voluntariamente las armas, no por estar ellos amenazados de destrucción, sino simplemente convenciendo a una gran mayoría de sus habitantes por medio de una publicidad y campaña de propaganda diseñada con mucha eficacia para ensalzar las virtudes de la unidad global. Si esto se hiciera con la intención de beneficiar a todo el mundo, por

utilizaría para formar efectivamente una Unidad. La conveniencia de este resultado depende, por supuesto, de la naturaleza de la Unidad que se creara y también de cómo se presentara el futuro de la vida inteligente en escenarios alternativos multipolares. Volveremos sobre estas preguntas en capítulos posteriores. Pero primero vamos a echar un vistazo más de cerca a por qué y en qué sentido una superinteligencia sería poderosa y eficaz logrando resultados en el mundo.

CAPÍTULO 6

Superpoderes cognitivos

de control que ilustre cómo un agente superinteligente, comenzando como mero software, podría establecerse como una Unidad. También ofreceremos algunas observaciones sobre la relación entre el poder sobre la naturaleza y el poder sobre otros agentes.

La razón principal de que la humanidad domine la Tierra es que nuestros cerebros tienen un conjunto ligeramente más amplio de facultades que otros animales.[1] Nuestra mayor inteligencia nos permite transmitir cultura de manera más eficiente, con el resultado de que el conocimiento y la tecnología se acumulan de una generación a la siguiente. Hasta el momento hemos acumulado suficiente cultura como para hacer posible el vuelo espacial, las bombas H, la ingeniería genética, la informática, las granjas industriales, los insecticidas, el movimiento internacional por la paz, y todos los accesorios de la civilización moderna. Los geólogos han comenzado a referirse a la época actual como el Antropoceno, en reconocimiento a la distintiva biótica, sedimentación y huellas geoquímicas de las actividades humanas.[2] Según una estimación, conformamos un 24% de la producción primaria neta del ecosistema planetario.[3] Y, sin embargo, estamos lejos de haber alcanzado los límites físicos de la tecnología.

la trayectoria de crecimiento de una IA seminal y sobre la psicología, las motivaciones y las capacidades de una superinteligencia madura.

Por ejemplo, una suposición común es que una máquina superinteligente sería como un ser humano muy inteligente, pero un poco friki. Imaginamos que la IA tendría una inteligencia libresca pero carecería de comprensión social, o que tendría una

asociación se fortalece cuando se observa que las personas que son buenas en el trabajo con las computadoras tienden ellos mismos a ser frikis. Así que es natural suponer que la inteligencia computacional más avanzada tendrá atributos similares, sólo que en un grado superior.

Esta heurística podría retener cierta validez en las primeras etapas de desarrollo de una IA seminal. (No hay razón alguna para suponer que se aplicaría a emulaciones o a humanos mejorados cognitivamente). En su etapa inmadura, lo que más tarde se convertiría en una IA superinteligente, podría aún carecer de muchas habilidades y talentos que los humanos tenemos de serie; y el patrón de fortalezas y debilidades una IA seminal *podría,* de hecho, tener alguna vaga semejanza con un friki de gran coeficiente intelectual. La característica más esencial de una IA seminal, además de ser fácil de mejorar (al tener una baja resistencia al progreso), es su capacidad para ejercer una potencia de optimización que amplifique la inteligencia de un sistema: una habilidad

que presumiblemente está muy relacionada con un buen desempeño en matemáticas, programación, ingeniería, investigación informática, y otras actividades "frikis". Sin embargo, incluso si una IA seminal tiene un perfil friki en una determinada etapa de su desarrollo, esto no implica que cuando se convierta en una superinteligencia madura esté igualmente limitada. Recordemos la distinción entre alcance directo e indirecto. Con suficiente habilidad de amplificación de la inteligencia, el resto de las capacidades intelectuales estarían al alcance indirecto de un sistema: el sistema podría desarrollar nuevos módulos cognitivos y habilidades según sea necesario, incluyendo la empatía, la perspicacia política, y cualesquiera otros poderes estereotipados que faltarían en personalidades computerizadas.

Aunque [illegible] que una superinteligencia podría tener todas las habili-

vimos en un capítulo anterior, ha sido particularmente enfático condenando este

adecuado pensar que una IA es inteligente como decimos que un ser humano promedio es inteligente en comparación con un escarabajo o un gusano.

Sería conveniente si pudiéramos cuantificar el calibre cognitivo de un sistema cognitivo cualquiera utilizando alguna métrica familiar, como las puntuaciones de CI o alguna versión de las valoraciones de Elo, que miden las capacidades relativas de los

dida que tenemos una idea de cómo se correlacionan con resultados relevantes en sentido práctico.[5] Por ejemplo, tenemos datos que muestran que las personas con un coeficiente intelectual de 130 tienen más probabilidades de sobresalir en la escuela y de tener éxito en una amplia gama de trabajos cognitivamente exigentes que los que tienen un coeficiente intelectual de 90. Pero supongamos que pudiéramos establecer de alguna manera que una determinada IA futura tuviera un coeficiente intelectual de 6.455: entonces, ¿qué? Realmente no tendríamos ni idea de lo que una IA como esa podría hacer. Ni siquiera sabríamos si tal IA tendría tanta inteligencia general como un adulto humano normal —quizás la IA tendría un conjunto de algoritmos muy especializados que le permitirían resolver preguntas típicas de las pruebas de inteligencia con eficacia sobrehumana, pero no mucho más.

Se han hecho algunos esfuerzos recientes para desarrollar mediciones de capacidad cognitiva que pudieran aplicarse a una gama más amplia de sistemas de proce-

dirección, si pudiera superar varias dificultades técnicas, podría llegar a ser bastante útil para algunos fines científicos, incluyendo el desarrollo de la IA. Para los fines de la presente investigación, sin embargo, su utilidad sería limitada, pues permaneceríamos ignorantes acerca de lo que implicaría una puntuación de rendimiento sobrehumano en relación a la capacidad real para lograr resultados prácticos importantes en el mundo.

Por tanto, servirá mejor a nuestros propósitos enumerar algunas tareas de importancia estratégica y caracterizar luego los sistemas cognitivos hipotéticos en función de cuáles tienen o carecen de las habilidades necesarias para tener éxito en estas tareas. Véase Tabla 8. Diremos que un sistema que destaca lo suficiente en cualquiera de

como la conveniencia prudencial de la asignación de recursos resultante. A estos beneficios de la colaboración puede ser posible añadir uno más: que una colaboración pre-transición más amplia podría ayudar con los problemas de coordinación importantes de la era post-transición.

idea de que la superinteligencia debería contribuir al bien común. Una norma de este tipo podría ser formulada como sigue:

El principio del bien común

La superinteligencia sólo debe desarrollarse para el beneficio de toda la humanidad y al servicio de ideales éticos ampliamente compartidos.[47]

cer la petición de compartir universalmente de los beneficios de la superinteligencia mediante la adopción de una "cláusula de bonanza" que implicaría que los beneficios hasta algún tope muy alto (por ejemplo, un billón de dólares anuales) se distribuirían de manera ordinaria entre los accionistas de la empresa y otros demandantes legales, y que sólo los beneficios que superaran ese umbral serían distribuidos a toda la humanidad de manera uniforme (o de otra manera de acuerdo a criterios morales universales). La adopción de unas "cláusulas de bonanza" de este tipo debería ser

tía valiosa (en la medida en que se pudiera confiar en tales compromisos) de que si alguna vez alguna empresa privada *diera* en el clavo con la explosión de inteligencia, todo el mundo compartiría la mayor parte de los beneficios. La misma idea se podría aplicar a entidades distintas de las empresas. Por ejemplo, los Estados podrían estar de acuerdo en que si alguna vez el PIB de cualquier Estado superara una fracción muy alta (por ejemplo, el 90%) del PIB mundial, el exceso debería distribuirse de manera

CAPÍTULO 1: DESARROLLOS DEL PASADO Y POSIBILIDADES DEL PRESENTE

1. A día de hoy la renta de subsistencia es de aproximadamente $ 400 (Chen y Ravallion, 2010). Un millón de ingresos de subsistencia son, pues, $ 400, 000,000. El producto bruto mundial actual es de aproximadamente $ 60.000.000.000.000 millones y en los últimos años ha crecido a un ritmo anual de alrededor del 4% (la tasa de crecimiento anual compuesta a partir de 1950, en base a Maddison [2010]). Estas cifras arrojan la estimación mencionada en el texto, que por supuesto es sólo una aproximación de cierto orden de magnitud. Si miramos directamente a las cifras de población, nos encontramos con que actualmente la población mundial necesita aproximadamente una semana y media para crecer en un millón; pero esto subestima el ritmo de crecimiento de la economía ya que el ingreso per cápita también estaría aumentando. En el 5000 a.C., tras la revolución agrícola, la población mundial estaba creciendo a un ritmo de alrededor de 1 millón por cada 200 años —una gran aceleración si lo comparamos con el ritmo de quizá un millón por cada millón años de principios de prehistoria humanoide— por lo que [illegible] una gran aceleración. Aun así, es impresionante que la cantidad de cre-

1958).

3. Hanson (2000).
4. Vinge (1993); Kurzweil (2005).
5. Sandberg (2010).
6. Van Zanden (2003); Maddison (1999, 2001); De Long (1998).
7. Dos declaraciones optimistas habitualmente repetidas en la década de 1960: "Las máquinas serán capaces, dentro de veinte años, de hacer cualquier trabajo que un hombre pueda hacer" (Simon, 1965, 96);

43. Esta tabla está tomada de Shulman y Bostrom (2014). Se basa en un modelo de juguete que asume una distribución gaussiana de CIs anticipados entre embriones con una desviación estándar de 7,5 puntos. La cantidad de mejora de la cognición que se reparta en los diferentes embriones depende de cómo los embriones se diferencian unos de los otros en las variantes genéticas aditivas cuyos efectos se

44. Con información imperfecta acerca de los efectos genéticos aditivos sobre la capacidad cognitiva, se reducirían los tamaños del efecto. Sin embargo, incluso una pequeña cantidad de conocimiento ayudaría mucho, porque las ganancias derivadas de la selección no escalan de manera lineal con la porción de variación que podemos predecir. En su lugar, la eficacia de nuestra selección depende de la predicción en la desviación estándar de CI medio, que escala como la raíz cuadrada de la variación. Por ejemplo, si se pudiera dar cuenta del 12,5% de la variación, esto podría ofrecer efectos medios tan grandes como los de la Tabla 1, que asumen el 50%. Para comparar, un estudio reciente (Rietveld et al., 2013) afirma que ya tienen identificado un 2,5% de variación.
45. En comparación, una práctica habitual hoy en día consiste en la creación de menos de diez embriones.

2007).

47. La opinión es de Katsuhiko Hayashi, según se informa en Cyranoski (2013). El Grupo Hinxton, un consorcio internacional de científicos que analiza la ética y los desafíos de las células madre, predijo en 2008 que los gametos derivados de células madre humanas estarían disponibles dentro de diez años (Grupo Hinxton, 2008), y los desarrollos hasta ahora son ampliamente consistentes con esto.
48. Sparrow (2013); Miller (2012); The Uncertain Future (2012).
49. Sparrow (2013).
50. Las preocupaciones seculares podrían centrarse en los impactos previstos en desigualdad social, la seguridad médica del procedimiento, los temores de una "carrera de ratas" mejorativa, los derechos y responsabilidades de los padres respecto de su descendencia, la sombra de la eugenesia del siglo XX, el

51. Para evitar los efectos negativos de la endogamia, la selección de embriones por iteración requeriría o bien un gran suministro de partida por parte de los donantes o el gasto de sustancial energía selectiva para reducir alelos recesivos perjudiciales. Cualquier alternativa tendería a empujar a que la descendencia tuviera una relación genética menos estrecha con sus padres (y estuvieran más relacionados entre sí).
52. Adaptado de Shulman y Bostrom (2014).
53. Bostrom (2008b).

generación. La mayoría de estas mutaciones no afectan al funcionamiento, o lo hacen sólo en un grado imperceptiblemente pequeño; pero los efectos combinados de muchas mutaciones perjudiciales muy pequeñas podrían implicar una importante pérdida de aptitud. Véase también Crow (2000).

59. Estos compuestos fueron creados por Mike Mike a partir de fotografías individuales tomadas por Virtual Flavius (Mike, 2013).
60. Éstas podrían, por supuesto, tener algunos efectos más rápidos —por ejemplo, cambiando las expectativas de la gente sobre lo que está por venir.
61. Louis Harris & Associates (1969); Mason (2003).
62. Kalfoglou et al. (2004).
63. Los datos son, obviamente, limitados, pero los individuos seleccionados para los resultados 1 entre 10.000 en las pruebas de habilidad de la infancia han demostrado, en estudios longitudinales, ser mucho más propensos a convertirse en profesores titulares, conseguir patentes y tener más éxito en los negocios

las ganancias de una vida y con resultados no financieros tales como la esperanza de vida, las tasas de divorcio y la probabilidad de abandonar la escuela [Deary, 2012]). Un desplazamiento hacia arriba de la distribución de la capacidad cognitiva tendría desproporcionadamente grandes efectos en las partes inferiores, sobre todo aumentando el número de superdotados y reduciendo el número de personas con retraso y dificultades de aprendizaje. Véase también Bostrom y Ord (2006) y Sandberg y Savulescu (2011).
64. Por ejemplo, Warwick (2002). Stephen Hawking llegó a sugerir que dar este paso podría ser necesario con el fin de mantenerse al día con los avances en inteligencia artificial: "Tenemos que desarrollar lo más rápidamente posible las tecnologías que hacen posible una conexión directa entre el cerebro y el ordenador, para que los cerebros artificiales contribuyan a la inteligencia humana en lugar de oponerse

65. Véase Lebedev y Nicolelis (2006); Birbaumer et al. (2006); Mak y Wolpaw (2009); y Nicolelis y Lebedev (2009). Una visión más personal sobre el problema de la mejora a través de implantes se puede encontrar en Chorost (2005, cap. 11).
66. Smeding et al. (2006).
67. Degnan et al. (2002).
68. Dagnelie (2012); Shannon (2012).
69. Perlmutter y Mink (2006); Lyons (2011).

77. Se ha sugerido que deberíamos considerar a las personas jurídicas (empresas, sindicatos, gobiernos, iglesias, etc.) como agentes inteligentes artificiales, entidades con sensores y efectores, capaces de representar conocimiento y realizar inferencias y tomar medidas (por ejemplo, Kuipers [2012]; cf. Huebner [2008] para una discusión sobre la posibilidad de que existan representaciones colectivas). Son claramente poderosas y ecológicamente exitosas, aunque sus capacidades y estados internos son diferentes de los de los humanos.

men o a los responsables de presuntos problemas entre la población en general.

80. Uno podría imaginar técnicas de neuroimagen que permitieran detectar firmas neurales motivadas por cognición. Sin detección de autoengaño, la detección de mentiras favorecería a individuos que creen sus propias mentiras. Mejores pruebas para los tests de auto-engaño también podrían ser utilizadas para entrenar la racionalidad y para el estudio de la efectividad de las intervenciones dirigidas a reducir el sesgo.
81. Bell y Gemmel (2009). Un primer ejemplo se encuentra en la obra de Deb Roy del MIT, quien registró cada momento de los tres primeros años de vida de su hijo. El análisis de estos datos audiovisuales están dando información sobre el desarrollo del lenguaje; ver Roy (2012).

como se puede ver teniendo en cuenta la diferencia en la velocidad y energía de los procesos relevantes del cerebro en comparación con el procesamiento de información más eficiente. La velocidad de la luz es más de un millón de veces mayor que la de la transmisión neuronal, los enlaces sinápticos disipan un

cuerpos de 7 mm del robot "tinkerbell" que se mueven 260 veces más rápido respecto de la velocidad normal (Hanson, 1994).
6. Hanson (2012).
7. La "inteligencia colectiva" no se refiere a la paralelización a pequeña escala de hardware de computación, sino a la paralelización a nivel de agentes autónomos inteligentes como los seres humanos. La implementación de una sola emulación en una máquina paralela masiva podría resultar en una superin-

superior y que cuando se llegó a un cierto umbral de acumulación de complejidad cultural fue algo mucho más sencillo (Powell et al., 2009).
10. ¿Qué pasa con internet? No parece haber supuesto un impulso de gran tamaño. Quizás lo haga eventualmente. Tomó siglos o milenios para que los otros ejemplos enumerados aquí revelaran todo su potencial.
11. Esto, obviamente, no pretende ser un experimento mental realista. Un planeta lo suficientemente grande como para sostener siete mil billones de organismos humanos con la tecnología actual podría implosionar, a menos que estuviera hecho de una materia muy ligera o estuviera hueco y fuera sostenido por la presión o por otros medios artificiales. (Una esfera de Dyson o un planeta-caparazón podrían ser

do aislado, aún podría ser parte de una inteligencia colectiva más grande de lo que parece a simple vista —es decir, la inteligencia colectiva incluye no sólo a la presente, sino a todas las generaciones ancestrales, un agregado que pueden funcionar como un sistema de procesamiento de información volcado hacia adelante

que actualmente carecemos de los cerebros necesarios.

18. Cf. Boswell (1917); véase también Walker (2002).
19. Esto ocurre principalmente en ráfagas cortas de subconjuntos de neuronas —la mayoría tienen tasas de disparo más tranquilas (Gray y McCormick, 1996; Steriade et al., 1998). Hay algunas neuronas ("neuronas castañeantes", también conocidas como células de "estallido rítmicamente rápido") que pueden alcanzar frecuencias de disparo de hasta 750 Hz, pero éstos parecen ser valores atípicamente extremos.
20. Feldman y Ballard (1982).
21. La velocidad de conducción depende del diámetro del axón (axones más gruesos son más rápidos) y de si el axón está mielinizado. Dentro del sistema nervioso central, los retardos de transmisión pueden variar desde menos de un milisegundo hasta 100 ms (Kandel et al., 2000). La transmisión en fibras

su lugar haber un intervalo durante el cual el sistema se vuelva gradualmente capaz de superar al equipo de investigación externo en un número creciente de tareas de desarrollo y mejoras del sistema.

2. En el último medio siglo, ha habido al menos un escenario que se ha creído podría llevar a acabar con el orden mundial existente en pocos minutos u horas: una guerra termonuclear global.
3. Esto sería coherente con la observación de que el efecto de Flynn —el aumento secular en las puntuaciones de CI en la mayoría de poblaciones a un ritmo de unos 3 puntos de CI por década en los últimos 60 años más o menos— parece haber cesado o incluso revertido en los últimos años en algunos países altamente desarrollados como el Reino Unido, Dinamarca y Noruega (Teasdale y Owen, 2008; Sundet et al., 2004). La causa del efecto Flynn en el pasado —y la cuestión de hasta qué punto representa algún beneficio real en inteligencia general o simplemente una mejor habilidad en la resolución de tests de

grande.

4. Bostrom y Roache (2011).
5. La terapia génica somática podría eliminar el retraso madurativo, pero es técnicamente mucho más difícil que las intervenciones en la línea germinal y tiene un potencial máximo más pequeño.
6. El promedio de crecimiento de la productividad de la economía mundial por año durante el período 1960-2000 fue de 4,3% (Isaksson, 2007). Sólo una parte de este crecimiento de la productividad se debe al aumento de la eficiencia organizacional. Algunas redes o procesos organizativos están mejorando claramente a un ritmo mucho más rápido.
7. La evolución del cerebro biológico estuvo sujeta a muchas restricciones y compensaciones que se rebajarían drásticamente si la mente se trasladara a un medio digital. Por ejemplo, el tamaño del cerebro

no a los 75-125 miles de millones de neuronas eran comunes. Estos normalmente se basaban en contar manualmente la densidad de células en pequeñas regiones representativas (Azevedo et al., 2009).

24. Whitehead (2003).
25. Los sistemas de procesamiento de información pueden muy probablemente utilizar procesos a escala molecular para la computación y almacenamiento de datos y llegar a un tamaño al menos planetario de extensión. Los límites físicos al cálculo establecidos por la mecánica cuántica, la relatividad general y la termodinámica están, sin embargo, mucho más allá de ese nivel de "cerebro-Júpiter" (Sandberg, 1999; Lloyd, 2000).
26. Stansberry y Kudritzki (2012). La electricidad utilizada en los centros de datos en todo el mundo as-

8. Yudkowsky (2008a, 326). Para un análisis más reciente, véase Yudkowsky (2013).
9. La imagen muestra la capacidad cognitiva como un parámetro de una sola dimensión, para no complicar el dibujo. Pero esto no es esencial para el punto que aquí se defiende. Uno podría, por ejemplo, representarse en su lugar un perfil de la capacidad cognitiva como una hipersuperficie en un espacio multidimensional.
10. Lin et al. (2012).
11. Se obtiene un cierto aumento de inteligencia colectiva simplemente aumentando el número de inteligencias constituyentes. Si se hace, por lo menos debería permitir un mejor rendimiento general en

29. Ver Landauer (1986). Este estudio se basa en estimaciones experimentales de las tasas de aprendizaje y olvido en los seres humanos. Teniendo en cuenta que el aprendizaje implícito podría hacer variar la estimación un poco. Si se asume una capacidad de almacenamiento de ~ 1 bit por sinapsis, se obtiene un límite superior de la capacidad memorística humana de alrededor de 10^{15} bits. Para una visión general de las diferentes estimaciones, véase el Apéndice A de Sandberg y Bostrom (2008).
30. El canal del ruido puede desencadenar potenciales de acción, y el ruido sináptico produce una variabilidad significativa en la intensidad de las señales transmitidas. Los sistemas nerviosos parecen haber evolucionado para hacer numerosas compensaciones entre la tolerancia al ruido y sus costes (masa, ta-

27. La IA que construyamos tampoco necesita ser capaz de encontrar el [illegible] que [illegible], podría razonar sobre lo que una definición implícita tan compleja implicaría (tal vez mirando a su entorno y siguiendo la misma clase de razonamiento que estábamos siguiendo).
28. Cf. Capítulos 9 y 11.
29. Por ejemplo, el MDMA puede aumentar temporalmente la empatía; la oxitocina puede aumentar temporalmente la confianza (Vollenweider y col., 1998; Bartz et al., 2011). Sin embargo, los efectos parecen

31. El problema también podría ser menos [illegible] que tuviera acceso a una vigilancia avanzada o a técnicas biomédicas para la manipulación psicológica, o que fuera lo suficientemente rica como para permitirse una proporción muy alta de profesionales de seguridad que vigilaran a la ciudadanía regular (y unos a otros).
32. Cf. Armstrong (2007) y Shulman (2010b).
33. Una pregunta abierta es hasta qué punto se necesita un supervisor de nivel n controlar no sólo a los supervisados de su nivel (n - 1), sino también a los de su nivel (n - 2), con el fin de saber qué nivel (n - 1) de agentes está haciendo su trabajo correctamente. Y para saber que el nivel (n - 1) de agentes ha logrado vigilar con éxito el nivel (n - 2) de agentes, ¿es necesario que el agente de nivel n monitorice también a los agentes de nivel (n - 3)?

36. Para que esta garantía sea efectiva, se debería implementar de buena fe. Esto descartaría ciertos tipos de manipulación de las facultades emocionales y la toma de decisiones de la emulación que de otro modo podrían ser utilizadas (por ejemplo) para instalar un miedo a ser detenido o para evitar que la emulación evaluara racionalmente sus opciones.
37. Véase, por ejemplo, Brinton (1965); Goldstone (1980, 2001). (El progreso de las ciencias sociales sobre estas cuestiones podría ser un buen regalo para los déspotas del mundo, que podrían utilizar modelos predictivos más precisos de malestar social para optimizar sus estrategias de control de la población y para cortar suavemente las insurgencias que brotaran con menos fuerza letal).
38. Cf. Bostrom (2011a, 2009b).
39. En el caso de un sistema completamente artificial, podría ser posible obtener algunas de las ventajas de una estructura institucional sin crear realmente subagentes distintos. Un sistema podría incorporar múltiples perspectivas en su proceso de decisión sin dotar a cada uno de esos puntos de vista con su

8. Más precisamente, deberíamos aplazar su opinión, excepto en aquellos temas en los que tenemos buenas razones para suponer que nuestras creencias son más precisas. Por ejemplo, podríamos saber más acerca de lo que estamos pensando en un momento particular de lo que sabría la superinteligencia si no fuera capaz de escanear el cerebro. Sin embargo, podríamos omitir esta calificación si asumiéramos que la superinteligencia tendría acceso a nuestras opiniones; podríamos entonces también dejar a la superinteligencia la tarea de juzgar cuando nuestras opiniones son dignas de confianza. (Pueden per-

10. Por ejemplo, David Lewis propuso una *teoría disposicional del valor*, que sostiene, más o menos, que algo X es un valor para A si y sólo si A querría X si A fuera perfectamente racional y conociera perfectamente X (Smith et al., 1989). Ideas similares se habían presentado anteriormente; véase, por ejemplo, Sen y Williams (1982), Railton (1986), y Sidgwick y Jones (2010). A lo largo de las líneas algo similares, una forma común de justificación filosófica, *el método del equilibrio reflexivo*, propone un proceso de adaptación mutua iterativa entre nuestras intuiciones sobre casos particulares, las normas generales que creemos que gobiernan estos casos, y los principios según los cuales pensamos que estos elementos deberían ser revisados, para lograr un sistema más coherente; véase, por ejemplo, Rawls (1971) y Goodman (1954).
11. Es de suponer que la intención aquí es que cuando la IA actúe para prevenir este tipo de desastres, debe

[illegible] de influencia sobre la humanidad en otros aspectos

dores hubieran hecho algo de [illegible]

15. Algunos grupos religiosos ponen un fuerte énfasis en la fe, en contraposición a la razón, a la que pueden llegar a considerar —incluso en su forma hipotéticamente más idealizada e incluso después de que se hubieran estudiado las escrituras, la revelación y la exégesis, con ardor y con mente abierta— como insuficiente para el logro de conocimientos espirituales esenciales. Aquellos que sostienen tales opiniones podrían no considerar la VCE como una guía óptima para la toma de decisiones (aunque podrían preferirla a varias otras guías imperfectas que podrían en realidad ser seguidas si se evitara el enfoque VCE).
16. Una IA que actuara como una fuerza latente de la naturaleza que regulara las interacciones humanas ha sido referida como un "Sysop," una especie de "sistema operativo" para la cuestión de la civilización humana. Véase Yudkowsky (2001).
17. "*Podría*", porque está *condicionado* con que la voluntad coherente extrapolada de la humanidad no de-

significado. Y en la medida en que las afirmaciones morales son "susceptibles de ser verdaderas" (es decir, tienen un carácter proposicional subyacente que les permite ser verdaderas o falsas), la superinteligencia debería ser capaz de averiguar qué afirmaciones de la forma "Agente *X* debe ahora Φ" son verdaderas. Por lo menos, debería superarnos en esta tarea. Una IA que careciera inicialmente de tal capacidad para la cognición moral debería ser capaz de adquirirla si tuviera el superpoder de amplificación de la inteligencia. Una forma en la que la IA podría hacer esto es mediante la ingeniería inversa del pensamiento moral del cerebro humano para luego implementar un proceso similar pero ejecutándolo más rápido, alimentándolo con información fáctica más precisa y así sucesivamente.

21. Ya que no estamos seguros de la metaética, está la cuestión de qué podría hacer la IA si las condiciones previas para la RM no fueran alcanzadas. Una opción sería establecer que la IA se apagara a sí misma si

apagado) se invocaría. También podríamos especificar lo que debería suceder si hubiera varias acciones posibles, cada una de los cuales fuera moralmente correcta. Por ejemplo, podríamos decir que en tales casos la IA debería realizar (una de) las medidas admisibles que la extrapolación colectiva de la humanidad habría favorecido. También podríamos estipular lo que debería suceder si la verdadera teoría moral no empleara términos como "moralmente correcto" en su vocabulario básico. Por ejemplo, una teoría consecuencialista podría sostener que algunas acciones son mejores que otras, pero que no existe un umbral particular, correspondiente a la noción de que una acción sea "moralmente correcta". Podríamos entonces decir que si esta teoría es correcta, la RM debe realizar una de las mejores acciones moralmente factibles, si la hubiera; o, si hubiera un número infinito de acciones factibles tales que para cualquier acción viable hubiera una mejor, entonces tal vez la RM podría elegir cualquiera que fuera al

25. Posiblemente podría incluso argumentarse que se trata de una condición de idoneidad para cualquier

RM. Supongamos que estipulamos que la IA debería actuar (hacer lo que sería moralmente correcto hacer) sólo si fuera moralmente correcto para sus creadores haber construido la IA en primer lugar; de lo contrario, debería apagarse. Es difícil ver cómo estaríamos cometiendo ningún mal moral grave al crear una IA de este tipo, ya que si hubiéramos hecho mal en crearla, la única consecuencia sería que se crearía una IA que inmediatamente se apagaría a sí misma, suponiendo que la IA no hubiera cometido ningún delito mental hasta ese momento. (Nosotros, sin embargo, podríamos haber actuado erróneamente —por ejemplo, al no haber aprovechado la oportunidad para construir algún otro tipo de IA en su lugar).

Una segunda cuestión es la supererrogación. Supongamos que hay muchas acciones que la IA pudiera tomar, cada una de las cuales sería moralmente correcta —en el sentido de ser *moralmente permisible*, para que algunas de ellas fueran moralmente mejores que las otras. Una opción es hacer que

permisible, podría proteger nuestros intereses más de lo que haría una IA que sólo hiciera lo moralmente mejor.

27. Cuando la IA evalúa la legitimidad moral de nuestro acto de creación de la IA, debería interpretar la permisibilidad en su sentido objetivo. En el sentido corriente de "moralmente permisible," un médico actúa de manera moralmente permisible cuando receta el medicamento que cree que va a curar a su paciente, incluso si el paciente, sin el conocimiento del médico, es alérgico a la droga y muere como resultado. Centrándose en la permisibilidad moral objetiva se aprovecha la posición epistémica presumiblemente superior de la IA.
28. Más directamente, depende de las *creencias* de la IA sobre qué teoría ética es verdadera (o, más precisamente, de su distribución de probabilidad entre las teorías éticas).
29. Puede ser difícil imaginar cómo de superlativamente maravillosas podrían ser estas vidas físicamente

entornos naturales. Por otra parte, la mayoría de las personas se preocupan por sus descendientes, y los privilegios especiales podían ser otorgados a los hijos y nietos de los contribuyentes. Más especulativamente, la superinteligencia podría ser capaz de crear simulaciones relativamente fieles de algunas personas pasadas —simulaciones que serían conscientes y que se asemejarían al original lo suficiente como para contar como una forma de supervivencia (de acuerdo con los criterios de algunas personas). Éste sería presumiblemente más fácil para las personas que hubieran entrado en suspensión criogénica; pero

[illegible] a partir de otros registros conservados como la correspondencia, publicaciones, materiales audiovisuales y documentos digitales, o los recuerdos personales de otros supervivientes. Una superinteligencia también podría pensar en algunas posibilidades que no se nos ocurrirían fácilmente a nosotros

[illegible]

IA. La aplicación podría permitir a los valores que nuestra voluntad coherente extrapolada querría que la IA quisiera depender de la teoría de la decisión de la IA. Si nuestros seres idealizados supieran que estaban haciendo especificaciones de valor para una IA que estaba usando un tipo particular de teoría de la decisión, podrían ajustar sus especificaciones de valor de tal forma que la IA se comportase con benevolencia a pesar de su errónea teoría de la decisión —de manera muy similar a como se pueden anular los efectos distorsionadores de una lente colocando otra lente frente de ella que distorsiona de forma opuesta.

40. Algunos sistemas epistemológicos pueden no tener un fundamento claro en un sentido holístico. En

[illegible]

observadores debería recibir un impulso de probabilidad proporcional a [illegible] un argumento en contra de este principio, véase el experimento mental del "Filósofo presuntuoso" en Bostrom (2002a). Para una defensa del principio, véase Olum (2002); y para una crítica de esa defensa, véase Bostrom y Ćirković (2003). Las creencias sobre el supuesto de auto-referenciación pueden afectar a diversas hipótesis empíricas de importancia estratégica potencialmente [illegible] por ejemplo, las consideraciones tales como el argumento del "día del juicio final" de Carter-Leslie, el argumento de simulación, y los argumentos de "gran filtro". Véase Bostrom (2002a, 2003a, 2008a); Carter (1983); Ćirković et al. (2010); Hanson (1998d); Leslie (1996); Tegmark y Bostrom (2005). Algo similar podría hacerse en relación a

gínese que la arena que se vierte en esta caja representa el esfuerzo de investigación. Cómo se vierta la arena determina dónde se amontona en la caja. Pero si siguiéramos vertiendo la arena, el espacio entero finalmente se llenaría.

2. Bostrom (2002b).
3. Ésta no es la perspectiva desde la cual se ha contemplado la política sobre la ciencia tradicionalmente. Harvey Averch describe la política sobre ciencia y tecnología en Estados Unidos entre 1945 y 1984 como algo centrado en debates sobre el nivel óptimo de inversión pública en empresas de ciencia y

[illegible]

mayor aumento de la prosperidad económica de la nación y en la fuerza militar. En estos cálculos, el progreso tecnológico siempre se supone que es bueno. Pero Averch también describe el surgimiento de

[illegible]

artificial. La nanotecnología (o la biología sintética, o cualquiera que fuera el reto menor al que nos enfrentáramos primero) podría entonces servir como taburete que nos ayudaría a ascender al nivel de capacidad necesaria para hacer frente al desafío de mayor nivel de la superinteligencia.

Tal argumento tendría que ser evaluado caso por caso. Por ejemplo, en el caso de la nanotecnología, tendrían que considerarse varias consecuencias posibles, como el aumento en el rendimiento del hardware de los sustratos computacionales nanofabricados; los efectos del capital físico barato de fabricación en el crecimiento económico; la proliferación de la tecnología de vigilancia sofisticada; la posibilidad de que un Unidad pudiera surgir a través de los efectos directos o indirectos de un gran

[illegible]

ficado, sobre la base de que no parece (a primera vista) tener consecuencias observacionales; pero véase, por ejemplo, Shoemaker (1969).

8. El nivel de preparación no se mide por la cantidad de esfuerzo invertido en las actividades de preparación, sino por lo adecuadamente configuradas que estén realmente las condiciones y por lo bien [illegible] posicionados que estén para elegir la acción apropiada los que vayan a tomar las decisiones clave.
9. El grado de confianza internacional durante el período previo a la explosión de inteligencia también podría ser un factor. Consideraremos esto en la sección de "Colaboración" más adelante en el capítulo.

BIBLIOGRAFÍA

Acemoglu, Daron. 2003. "Labor- and Capital-Augmenting Technical Change." *Journal of the European Economic Association* 1 (1): 1–37.

Albertson, D. G., and Thomson, J. N. 1976. "The Pharynx of *Caenorhabditis Elegans*." *Philosophical Transactions of the Royal Society B: Biological Sciences* 275 (938): 299-325.

Allen, Robert C. 2008. "A Review of Gregory Clark's *A Farewell to Alms: A Brief Economic History of the World*." *Journal of Economic Literature* 46 (4): 946–73.

American Horse Council. 2005. "National Economic Impact of the US Horse Industry." Retrieved July 30, 2013. Available at http://www.horsecouncil.org/national-economic-impact-us-horse-industry.

Anand, Paul, Pattanaik, Prasanta, and Puppe, Clemens, eds. 2009. *The Oxford Handbook of Rational and Social Choice*. New York: Oxford University Press.

Andres, B., Koethe, U., Kroeger, T., Helmstaedter, M., Briggman, K. L., Denk, W., and Hamprecht, F. A. 2012. "3D Segmentation of SBFSEM Images of Neuropil by a Graphical Model over Supervoxel Boundaries." *Medical Image Analysis* 16 (4): 796–805.

Armstrong, Alex. 2012. "Computer Competes in Crossword Tournament." *I Programmer*, March 19.

Armstrong, Stuart. 2007. "Chaining God: A Qualitative Approach to AI, Trust and Moral Systems." Unpublished manuscript, October 20. Retrieved December 31, 2012. Available at http://www.neweuropeancentury.org/GodAI.pdf.

Armstrong, Stuart. 2010. *Utility Indifference*, Technical Report 2010-1. Oxford: Future of Humanity Institute, University of Oxford.

Armstrong, Stuart. 2013. "General Purpose Intelligence: Arguing the Orthogonality Thesis." *Analysis and Metaphysics* 12: 68–84.

Armstrong, Stuart, and Sandberg, Anders. 2013. "Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox." *Acta Astronautica* 89: 1–13.

Armstrong, Stuart, and Sotala, Kaj. 2012. "How We're Predicting AI—or Failing To." In *Beyond AI: Artificial Dreams*, edited by Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, and Radek Schuster, 52–75. Pilsen: University of West Bohemia. Retrieved February 2, 2013.

Asimov, Isaac. 1942. "Runaround." *Astounding Science-Fiction*, March, 94–103.

Asimov, Isaac. 1985. *Robots and Empire*. New York: Doubleday.

Aumann, Robert J. 1976. "Agreeing to Disagree." *Annals of Statistics* 4 (6): 1236–9.

Averch, Harvey Allen. 1985. *A Strategic Analysis of Science and Technology Policy*. Baltimore: Johns Hopkins University Press.

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., Jacob, W., Lent, R., and Herculano-Houzel, S. 2009. "Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-up Primate Brain." *Journal of Comparative Neurology* 513 (5): 532-41.

Baars, Bernard J. 1997. *In the Theater of Consciousness: The Workspace of the Mind*. New York: Oxford University Press.

Baratta, Joseph Preston. 2004. *The Politics of World Federation: United Nations, UN Reform, Atomic Control.* Westport, CT: Praeger.

Barber, E. J. W. 1991. *Prehistoric Textiles: The Development of Cloth in the Neolithic and Bronze Ages with Special Reference to the Aegean.* Princeton, NJ: Princeton University Press.

Bartels, J., Andreasen, D., Ehirim, P., Mao, H., Seibert, S., Wright, E. J., and Kennedy, P. 2008. "Neurotrophic Electrode: Method of Assembly and Implantation into Human Motor Speech Cortex." *Journal of Neuroscience Methods* 174 (2): 168-76.

Bartz, Jennifer A., Zaki, Jamil, Bolger, Niall, and Ochsner, Kevin N. 2011. "Social Effects of Oxytocin in Humans: Context and Person Matter." *Trends in Cognitive Science* 15 (7): 301-9.

Basten, Stuart, Lutz, Wolfgang, and Scherbov, Sergei. 2013. "Very Long Range Global Population Scenarios to 2300 and the Implications of Sustained Low Fertility." *Demographic Research* 28: 1145-66.

Baum, Eric B. 2004. *What Is Thought?* Bradford Books. Cambridge, MA: MIT Press.

Baum, Seth D., Goertzel, Ben, and Goertzel, Ted G. 2011. "How Long Until Human-Level AI? Results from an Expert Assessment." *Technological Forecasting and Social Change* 78 (1): 185-95.

Beal, J., and Winston, P. 2009. "Guest Editors' Introduction: The New Frontier of Human-Level Artificial Intelligence." *IEEE Intelligent Systems* 24 (4): 21-3.

Bell, C. Gordon, and Gemmell, Jim. 2009. *Total Recall: How the E-Memory Revolution Will Change Everything.*

Bostrom, Nick. 2003b. "Astronomical Waste: The Opportunity Cost of Delayed Technological Development." *Utilitas* 15 (3): 308-314.

Bostrom, Nick. 2003c. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence,* edited by Iva Smit and George E. Lasker, 2: 12-17. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

Bostrom, Nick. 2004. "The Future of Human Evolution." In *Two Hundred Years After Kant, Fifty Years After Turing,* edited by Charles Tandy, 2: 339-371. Death and Anti-Death. Palo Alto, CA: Ria University Press.

Bostrom, Nick. 2006a. "How Long Before Superintelligence?" *Linguistic and Philosophical Investigations* 5(1): 11-30.

Bostrom, Nick. 2006b. "Quantity of Experience: Brain-Duplication and Degrees of Consciousness." *Minds and Machines* 16 (2): 185-200.

Bostrom, Nick. 2006c. "What is a Singleton?" *Linguistic and Philosophical Investigations* 5 (2): 48-54.

Bostrom, Nick. 2007. "Technological Revolutions: Ethics and Policy in the Dark." In *Nanoscale: Issues and Perspectives for the Nano Century,* edited by Nigel M. de S. Cameron and M. Ellen Mitchell, 129-52. Hoboken, NJ: Wiley.

Bostrom, Nick. 2008a. "Where Are They? Why I Hope the Search for Extraterrestrial Life Finds Nothing."

ing 20 (2): 198-211.

Berliner, Hans J. 1980a. "Backgammon Computer-Program Beats World Champion." *Artificial Intelligence* 14 (2): 205-220.

Berliner, Hans J. 1980b. "Backgammon Program Beats World Champ." *SIGART Newsletter* 69: 6-9.

Bernardo, José M., and Smith, Adrian F. M. 1994. *Bayesian Theory,* 1st ed. Wiley Series in Probability & Statistics. New York: Wiley.

Birbaumer, N., Murguialday, A. R., and Cohen, L. 2008. "Brain-Computer Interface in Paralysis." *Current Opinion in Neurology* 21 (6): 634-8.

Bird, Jon, and Layzell, Paul. 2002. "The Evolved Radio and Its Implications for Modelling the Evolution of

Bostrom, Nick. 2011b. "Infinite Ethics." *Analysis and Metaphysics* 10: 9-59.

Bostrom, Nick. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." In "Theory and Philosophy of AI," edited by Vincent C. Müller, special issue, *Minds and Machines* 22 (2): 71-85.

Bostrom, Nick, and Ćirković, Milan M. 2003. "The Doomsday Argument and the Self-Indication Assumption: Reply to Olum." *Philosophical Quarterly* 53 (210): 83-91.

Bostrom, Nick, and Ord, Toby. 2006. "The Reversal Test: Eliminating the Status Quo Bias in Applied Ethics." *Ethics* 116 (4): 656-79.

Bostrom, Nick, and Roache, Rebecca. 2011. "Smart Policy: Cognitive Enhancement and the Public Interest." In *Enhancing Human Capacities,* edited by Julian Savulescu, Ruud ter Meulen, and Guy Kahane, 138-49.

Bouchard, T. J. 2004. "Genetic Influence on Human Psychological Traits: A Survey." *Current Directions in Psychological Science* 13 (4): 148-51.

Bourget, David, and Chalmers, David. 2009. "The PhilPapers Surveys." November. Available at http://philpapers.org/surveys/.

Bradbury, Robert J. 1999. "Matrioshka Brains." Archived version. As revised August 16, 2004. Available at

Burch-Brown, J. 2014. "Clues for Consequentialists." Utilitas 26 (1): 105-19.

Burke, Colin. 2001. "Agnes Meyer Driscoll vs. the Enigma and the Bombe." Unpublished manuscript. Retrieved February 22, 2015. Available at http://

Carmena, J. M., Lebedev, M. A., Crist, R. E., O'Doherty, J. E., Santucci, D. M., Henriquez, C. S., and Nicolelis, M. A. 2005. "Learning to Control a Brain-Machine Interface for Reaching and Grasping by Primates." Collaborative Biology Research Laboratory, Duke University.

Carroll, Bradley W., Ostrowski, Dale A. 2007. An Introduction to Modern Astrophysics. Pearson Addison Wesley.

Carroll, John E. 1996. Human Cognitive Abilities. Fourth Edition. Cambridge University Press.

Chason, R. J., Csokmay, J., Segars, J. H., DeCherney, A. H., and Armstrong, A. 2011. "Genetic Effects Upon Preimplantation Embryo Metabolism and Development." *Trends in Endocrinology*

Cirasella, Jill, and Kopec, Danny. 2006. "The History of Computer Games." Exhibit at Dartmouth Artificial Intelligence Conference: The Next Fifty Years (AI@50), Dartmouth College, July 13-15.

Ćirković, Milan M. 2004. "Forecast for the Next Eon: Applied Cosmology and the Long-Term Fate of Intelligent Beings." *Foundations of Physics* 34 (2): 239-61.

Ćirković, Milan M., Sandberg, Anders, and Bostrom, Nick. 2010. "Anthropic Shadow: Observation Selection

Coase, R. H. 1937. "The Nature of the Firm." Economica 4 (16): 386-405.

Cochran, Gregory, and Harpending, Henry. 2009. *The 10,000 Year Explosion: How Civilization Accelerated Human Evolution*. New York: Basic Books.

Cook, James Gordon. 1984. *Handbook of Textile Fibres: Natural Fibres*. Cambridge: Woodhead.

Dai, Wei. 2009. "Towards a New Decision Theory." *Less Wrong* (blog), August 13.

Dalrymple, David. 2011. "Comment on Kaufman, J. 'Whole Brain Emulation: Looking at Progress on C. Ele-

De Raedt, Luc, and Flach, Peter, eds. 2001. *Machine Learning: ECML 2001: 12th European Conference on Machine Learning, Freiburg, Germany, September 5-7, 2001. Proceedings.* Lecture Notes in Computer Science 2167. New York: Springer.

Dean, Cornelia. 2005. "Scientific Savvy? In U.S., Not Much." *New York Times*, August 30.

Deary, Ian J. 2001. "Human Intelligence Differences: A Recent History." *Trends in Cognitive Sciences* 5 (3): 127-30.

Deary, Ian J. 2012. "Intelligence." *Annual Review of Psychology* 63: 453-82.

Deary, Ian J., Penke, L., and Johnson, W. 2010. "The Neuroscience of Human Intelligence Differences." *Nature Reviews Neuroscience* 11 (3): 201-11.

Degnan, G. G., Wind, T. C., Jones, E. V., and Edlich, R. F. 2002. "Functional Electrical Stimulation in Tetraplegic Patients to Restore Hand Function." *Journal of Long-Term Effects of Medical Implants* 12 (3): 175-88.

Deary, I. J., Daniels, M., and Pooder, K. 1997. "The Heritability of IQ." *Nature* 388 (6641): 468-71.

Drescher, Gary L. 2006. *Good and Real: Demystifying Paradoxes from Physics to Ethics.* Bradford Books. Cambridge, MA: MIT Press.

Drexler, K. Eric. 1986. *Engines of Creation.* Garden City, NY: Anchor.

Drexler, K. Eric. 1992. *Nanosystems: Molecular Machinery, Manufacturing, and Computation.* New York: Wiley.

Drexler, K. Eric. 2013. *Radical Abundance: How a Revolution in Nanotechnology Will Change Civilization.* New York: PublicAffairs.

Driscoll, Kevin. 2012. "Code Critique: 'Altair Music of a Sort.'" Paper presented at Critical Code Studies Working Group Online Conference, 2012, February 6.

Dyson, Freeman J. 1960. "Search for Artificial Stellar Sources of Infrared Radiation." *Science* 131 (3414): 1667-1668.

Dyson, Freeman J. 1979. *Disturbing the Universe.* 1st ed. Sloan Foundation Science Series. New York: Harper & Row.

Elga, Adam. 2004. "Defeating Dr. Evil with Self-Locating Belief." *Philosophy and Phenomenological Research*

Faisal, A. A., White, J. A., and Laughlin, S. B. 2005. "Ion-Channel Noise Places Limits on the Miniaturization of the Brain's Wiring." *Current Biology* 15 (12): 1143-9.

Feldman, Jacob. 2000. "Minimization of Boolean Complexity in Human Concept Learning." *Nature* 407 (6804): 630-3.

Feldman, J. A., and Ballard, Dana H. 1982. "Connectionist Models and Their Properties." *Cognitive Science* 6 (3): 205-254.

Foley, J. A., Monfreda, C., Ramankutty, N., and Zaks, D. 2007. "Our Share of the Planetary Pie." *Proceedings of the National Academy of Sciences of the United States of America* 104 (31): 12585-6.

Forgas, Joseph P., Cooper, Joel, and Crano, William D., eds. 2010. *The Psychology of Attitudes and Attitude Change.* Sydney Symposium of Social Psychology. New York: Psychology Press.

Frank, Robert H. 1999. *Luxury Fever: Why Money Fails to Satisfy in an Era of Excess.* New York: Free Press.

Fredriksen, Kaja Bonesmo. 2012. *Less Income Inequality and More Growth - Are They Compatible?: Part 6. The*

Freitas, Robert A., Jr., and Merkle, Ralph C. 2004. *Kinematic Self-Replicating Machines.* Georgetown, TX: Landes Bioscience.

Gaddis, John Lewis. 1982. *Strategies of Containment: A Critical Appraisal of Postwar American National Security Policy.* New York: Oxford University Press.

Gammoned.net. 2012. "Snowie." Archived version. Retrieved June 30. Available at http://web.archive.org/web/20070920191840/http://www.gammoned.com/snowie.html.

Gates, Bill. 1975. "Software Contest Winners Announced." *Computer Notes* 1 (2): 1.

Georgieff, Michael K. 2007. "Nutrition and the Developing Brain: Nutrient Priorities and Measurement." *American Journal of Clinical Nutrition* 85 (2): 614S-620S.

Gianaroli, Luca. 2000. "Preimplantation Genetic Diagnosis: Polar Body and Embryo Biopsy." Supplement, *Human Reproduction* 15 (4): 69-75.

Gilovich, Thomas, Griffin, Dale, and Kahneman, Daniel, eds. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment.* New York: Cambridge University Press.

Hanson, Robin. 1998d. "Must Early Life Be Easy? The Rhythm of Major Evolutionary Transitions." Unpublished manuscript, September 23. Retrieved August 12, 2012. Available at http://hanson.gmu.edu/hardstep.pdf.

Hanson, Robin. 2000. "Shall We Vote on Values, But Bet on Beliefs?" Unpublished manuscript, September. Last revised October 2007. Available at http://hanson.gmu.edu/futarchy.pdf.

Hanson, Robin. 2006. "Uncommon Priors Require Origin Disputes." *Theory and Decision* 61 (4): 319-328.

bility and Sequential Decisions." In De Raedt and Flach 2001, 226-38.

Hutter, Marcus. 2005. *Universal Artificial Intelligencet: Sequential Decisions Based On Algorithmic Probability.* Texts in Theoretical Computer Science. Berlin: Springer.

Iliadou, A. N., Janson, P. C., and Cnattingius, S. 2011. "Epigenetics and Assisted Reproductive Technology." *Journal of Internal Medicine* 270 (5): 414-20.

O'Neill, Gerard K. 1974. "The Colonization of Space." *Physics Today* 27 (9): 32-40.

Pasquinelli, Elena, Federici, Stefano, and Belardinelli, Marta Olivetti. 2012. "Cognitive Prerequisites for the Use of External Tools in Autism Spectrum Disorders and Down Syndrome." *Research in Developmental Disabilities* 33 (1): 214-23.

Pasquinelli, Elena, Federici, Stefano, and Belardinelli, Marta Olivetti. 2012. "Open Questions in Cognitive Research: Neuroscience, Computation and Technology." *Technology* 7 (2): 89-105.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York: Cambridge University Press.

Perlmutter, J. S., and Mink, J. W. 2006. "Deep Brain Stimulation." *Annual Review of Neuroscience* 29: 229-57.

Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. New York: Viking.

Plomin, R., Haworth, C. M., Meaburn, E. L., Price, T. S., Wellcome Trust Case Control Consortium 2, and Davis, O. S. 2013. "Common DNA Markers Can Account for More than Half of the Genetic Influence on Cognitive Abilities." *Psychological Science* 24 (2): 562-8.

Popper, Nathaniel. 2012. "Flood of Errant Trades Is a Black Eye for Wall Street." *New York Times*, August 1.

Pourret, Olivier, Naim, Patrick, and Marcot, Bruce, eds. 2008. *Bayesian Networks: A Practical Guide to Appli*

Quine, Willard Van Orman, and Ullian, Joseph Silbert. 1970. *The Web of Belief*.

Railton, Peter. 1986. "Facts and Values." *Philosophical Topics* 14 (2): 5-31.

Rajab, Maheeb Abu, Zarfoss, Jay, Monrose, Fabian, and Terzis, Andreas. 2006. "A Multifaceted Approach to Understanding the Botnet Phenomenon." In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, 41-52. New York: ACM.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Belknap.

Read, J. I., and Trentham, Neil. 2005. "The Baryonic Mass Function of Galaxies." *Philosophical*

Richardson, Robert, and Bostrom, Nick. 2009. "The Fallacy of the Fallacy of Quantum Consciousness." *Philosophical*

Roache, Rebecca. 2008. "Ethics, Speculation, and Values." *NanoEthics* 2 (3): 317-27.

Russell, Bertrand. 1986. "The Philosophy of Logical Atomism." In *The Philosophy of Logical Atomism and Other Essays 1914-1919*, edited by John G. Slater, 8: 157-244. The Collected Papers of Bertrand Russell. Boston: Allen & Unwin.

Russell, Bertrand, and Griffin, Nicholas. 2001. *The Selected Letters of Bertrand Russell: The Public Years, 1914-1970*. New York: Routledge.

Russell, Stuart J., and Norvig, Peter. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Sabrosky, Curtis W. 1952. "How Many Insects Are There?" In *Insects*, edited by United States Department of Agriculture, 1-7. Yearbook of Agriculture. Washington, DC: United States Government Printing Office.

Sandberg, Anders. 2010. "When Software Goes Mental: Why Artificial Minds Mean Fast Endogenous Growth."

Sandberg, Anders. 2013. "An Overview of Models of Technological Singularity."

Sandberg, Anders. 2013. "Feasibility of Whole Brain Emulation."

Sandberg, Anders, and Bostrom, Nick. 2008. "Whole Brain Emulation: A Roadmap." Technical Report 2008-3. Future of Humanity Institute, Oxford University.

Sandberg, Anders, and Bostrom, Nick. 2011. "Machine Intelligence Survey." Technical Report 2011-1. Future of Humanity Institute, Oxford University.

ÍNDICE

A

acelerador de desarrollo macro-estructural 235, 236
acelerador universal 234
Acumulación de valores 189-190, 208
adquisición de recursos 113-114, 123, 153, 193
[illegible]
[illegible]
algoritmos genéticos 9, 15, 24, 27, 238, 241
Andamiaje motivacional 191-192
anticipación de consecuencias (argumentos de) 240
antrópica 127, 134, 143, 157-158, 222, 225, 285
aprendizaje artificial 8-9, 11-13, 15-16, 18, 28, 188
Aprendizaje de valores 192, 208
aprendizaje por refuerzo 12
Arendt, Hannah 105
Armstrong, Stuart 263, 265-267, 280, 282, 286, 293, 296, 303
Asimov, Isaac 139, 286
atraco Pascaliano 223, 300
atrofia 129, 135-137, 143
aumentación 138, 142-143, 201, 203
[illegible]

B

[illegible]

C

C. elegans 34-35, 268
cables trampa 129, 136-138, 143
capacidad de memoria 7, 72
capital 39, 48, 68, 84, 88, 99, 161-163, 165-167, 170
carga mutacional 41
carrera tecnológica 80-81, 205, 232-233, 248
Carta blanca (juego) 13
Catástrofe de la hipótesis de
CHINOOK 12
Christiano, Paul 200, 208, 295
cláusula de bonanza 255, 305
clonación 41
definición 41
Código alemán Enigma 87
cognición biológica 22, 36, 44, 51, 233
[illegible]
193, 219
[illegible]
222
control de la capacidad 127, 129, 135, 143, 158, 185
Copérnico, Nicolás 14
costes de negociación 182
crecimiento de 1, 27, 49, 74, 79, 92, 164-165, 263, 275, 283
crecimiento económico 163-164, 233, 263, 276, 301
crecimiento moral 214
Crimen mental 125
criptografía 82