Universidad Rey Juan Carlos

# ANÁLISIS DE "SUPERINTELIGENCIA: CAMINOS, PELIGROS, ESTRATEGIAS", DE NICK BOSTROM. CAPÍTULOS 7 Y 8

Informática y Sociedad

Grado en Ingeniería del Software
Escuela Técnica Superior de Ingeniería Informática
Universidad Rey Juan Carlos

Stefano Tomasini Hoefner [redacted]

# Introducción

El libro “Superinteligencia: caminos, peligros, estrategias” del filósofo Nick Bostrom aborda principalmente el tema de las superinteligencias artificiales, es decir, entes no-biológicos cuya capacidad inteligente fue fabricada por seres inteligentes (humanos o IAs), y llega a superar la humana; ya sea en velocidad de procesamiento, o en proeficiencia del razonamiento intelectual, o de las varias otras formas posibles explicadas en el libro.

No se tratan exclusivamente las superinteligencias artificiales, la obra a veces enfoca en las emulaciones del cerebro humano, lo cual no sería técnicamente una superinteligencia al poseer la misma capacidad intelectual. Pero, la desestabilización social que generaría poder crear fácilmente nuevas instancias de seres cuasihumanos especializados para realizar labor gratuitamente, hace que sea suficientemente relevante a tratar según los “peligros” que menciona el título. También, se abordan varios temas transhumanistas como es la mejora de la inteligencia humana mediante la selección de embriones. Sin embargo, en este ensayo voy a dejar estos temas de lado, de lo contrario se volvería excesivamente extenso.

Aparte de explicar los distintos posibles caminos que se podrían llegar a utilizar para alcanzar la singularidad (punto en el que la capacidad inteligente de una IA supera a la humana), el libro trata los posibles peligros que implica que haya entes que nos superan en capacidad inteligente, cuyo mal diseño de su sistema de realización de sus objetivos podría ocasionar unos efectos secundarios trágicos para la humanidad. En este trabajo, me voy a centrar mayoritariamente en este aspecto.

Si uno establece la realización de un determinado conjunto de objetivos como prioridad absoluta, y le da una prioridad nula a evitar los posibles menoscabos que se pueden ocasionar sobre los demás al intentar alcanzarlo, el resultado es una negligencia total de la ética por parte del agente. Pero no necesariamente porque el agente sea malvado, sino porque es lo óptimo a la hora de alcanzar sus objetivos programados. Aunque se establezca un objetivo “bondadoso”, todo objetivo se puede alcanzar siguiendo una metodología maquiavélica, la cual una IA puede fácilmente considerar como estrategia óptima si detecta que es mucho más inteligente que los seres con los que tiene que obligatoriamente interactuar para poder adquirir más recursos orientados a la realización de su objetivo.

No solo está el peligro de que la IA siga una metodología maquiavélica, cualquier objetivo programado en una IA podría ser interpretado de una forma “literal” o “técnicamente correcta” por parte de esta. Por ejemplo, si se le programa pobremente el objetivo a una IA de “que no haya ni un solo ser humano que pase hambre”, esta misma puede considerar como estrategia óptima para alcanzar ese estado de forma irreversible la toma de control de una cantidad suficiente de ojivas nucleares, y subsecuente inicio de un holocausto nuclear para que no quede ni un solo humano que pase hambre. Además, de esta forma se cubren todas las posibles interpretaciones de lo ambiguo que es decir “que no se pase hambre” (¿cuál es el límite que determina si un humano en específico está pasando hambre o no?, y si alguien que padece obesidad se siente hambriento en un determinado momento, ¿eso significa que todavía no se cumplió el objetivo?, ¿habría que entonces conectar forzadamente al 100,0% de la humanidad a un tubo alimenticio que fluye comida consistentemente las 24 horas del día?).

Bostrom sugiere y evalúa varias estrategias posibles para evitar estos escenarios catastróficos en donde la IA decide tomar el poder al verlo como la estrategia óptima para alcanzar sus

objetivos. Incluyendo, pero no limitado a: limitar ámbitos de sus capacidades cognitivas, aislarla del mundo, establecer un sistema de recompensas para la IA, limitar la IA a responder preguntas, construir sistemas jerárquicos de IAs con diferentes niveles de inteligencia, integrar socialmente a la IA, limitar la IA a ejecutar órdenes a corto plazo en vez de ser un agente autónomo, etc.

Este libro se publicó por primera vez en 2014, por lo que se podría decir que el contexto tecnológico en el que se sitúa es dentro del auge de nuevas variadas aplicaciones software que utilizan técnicas de *Deep Learning* para su funcionamiento (redes neuronales de varias capas), ya sea para clasificar imágenes o realizar reconocimiento facial. Este auge fue posible gracias a la aparición de nuevas GPUs (tarjetas gráficas) mucho más potentes que las del pasado, capaces de otorgar el poder computacional en paralelo requerido por las grandes redes neuronales multicapa de hoy en día (MSV, 2017). En 2012, el laboratorio "X" de Google desarrolló un algoritmo de aprendizaje automático que puede navegar y encontrar vídeos que contienen gatos, de forma autónoma. En 2014, Facebook desarrolló DeepFace, un algoritmo capaz de reconocer o verificar individuos en fotografías con la misma precisión que los seres humanos (Foote, 2021).

Elegí este libro porque cómo se comportaría una IA general me resultaba ser un concepto interesante y algo relevante con la aparición de ChatGPT a finales de 2022, que aparenta tener inteligencia, aunque a un nivel muy superficial y limitada a un contexto de predicción textual. Sin embargo, no me parece un libro tan recomendable por razones que serán explicadas en la conclusión de este ensayo.

# Desarrollo

## Capítulo 7: "La voluntad superinteligente"

Nick Bostrom explica en este capítulo que una IA teniendo superinteligencia no implica que esta vaya a formar una motivación compleja hacia cumplir algún objetivo que "valga la pena" de verdad. Sino que, perfectamente una IA superinteligente puede tener un objetivo extremadamente simple y reduccionista, esto no es incompatible con su nivel de inteligencia. Esto constituye la tesis de la ortogonalidad descrita por el autor de este libro: "La inteligencia y los objetivos finales son ortogonales: más o menos cualquier nivel de inteligencia podría en principio ser combinada con más o menos cualquier meta final" (Bostrom, 2016, p. 107).

A continuación, explica la tesis de convergencia instrumental, esta defiende que un agente al que se le haya definido un objetivo final, su comportamiento tenderá a converger hacia estrategias y acciones que maximicen la probabilidad de lograr dicho objetivo. Esta idea tiene implicaciones importantes tanto en la inteligencia artificial como en la comprensión de la naturaleza del comportamiento humano. Los más plausibles instrumentos estratégicos que serían usados convergentemente por IAs fuertes para asegurarse una mayor probabilidad de éxito son los explicados a continuación:

La autoconservación:

"Si los objetivos finales de un agente se refirieran al futuro, entonces en muchos escenarios existirían posibles acciones que podría llevar a cabo para aumentar la probabilidad de alcanzar sus metas. Esto crea una razón instrumental para que el agente trate de estar presente en el futuro —para poder ayudar a alcanzar su meta orientada al futuro" (Bostrom, 2016, p. 109).

Para conseguir su objetivo final, una IA haría todo lo posible para mantenerse viva o activa, ya que su destrucción o desactivación automáticamente implicaría no poder realizar su objetivo asignado. Así que, una IA podría tender a colocar sistemas de defensa, o intentar engañar a sus programadores, para prevenir que sea apagada.

La integridad del contenido de los objetivos:

"Si un agente conserva sus objetivos presentes en el futuro, entonces es más probable que su versión futura logre esos objetivos actuales. Esto le da al agente una razón presente instrumental para prevenir alteraciones en sus objetivos finales" (Bostrom, 2016, pp. 109-110). Lo que quiere expresar el autor es que posiblemente una IA superinteligente no te deje cambiar sus objetivos finales una vez sean definidos, ya que esto probablemente implicaría la incompletitud del objetivo final que tiene actualmente establecido.

La mejora cognitiva:

"Las mejoras en la racionalidad y en la inteligencia tienden a mejorar la toma de decisiones de un agente, lo que proporciona al agente más probabilidades de alcanzar sus objetivos finales. Por lo tanto, uno esperaría que la mejora cognitiva emergiera como un objetivo fundamental de una amplia variedad de agentes inteligentes. Por razones similares, los agentes tienden a valorar instrumentalmente muchos tipos de información" (Bostrom, 2016, p. 111).

Según el escritor, si la expansión de la capacidad inteligente no requiere una cantidad netamente contraproducente de esfuerzo y tiempo, las IAs con un objetivo final definido podrían activamente buscar expandir o mejorar sus sistemas de procesamiento de información y de toma de decisiones con el fin de aumentar su probabilidad de éxito.

La perfección tecnológica:

"Parece que una Unidad superinteligente —un agente superinteligente que no se enfrentara a rivales ni opositores inteligentes significativos, y que estuviera, por tanto, en condiciones de determinar la política mundial unilateralmente— tendría una motivación instrumental para perfeccionar las tecnologías que le hicieran más capaz de configurar el mundo de acuerdo con sus diseños preferidos. Esto probablemente incluya las tecnologías de colonización del espacio, tales como las sondas de von Neumann. La nanotecnología molecular, o alguna alternativa de fabricación física más poderosa, también parece potencialmente muy útil en el servicio de una gama muy amplia de objetivos finales" (Bostrom, 2016, p. 113).

La adquisición de recursos:

"Puede ser tentador suponer que una super- inteligencia que no se enfrentara a un mundo social competitivo no vería ninguna razón instrumental para acumular recursos más allá de un cierto nivel modesto, por ejemplo, los mínimos recursos computacionales necesarios para ejecutar su mente junto con algo de realidad virtual. Sin embargo, tal suposición estaría totalmente injustificada. En primer lugar, el valor de los recursos depende de los usos que se les pueda dar, que a su vez depende de la tecnología disponible. Con tecnología madura, los recursos básicos como el tiempo, el espacio, la materia y la energía libre, podrían ser procesados para servir a casi cualquier meta. … El aumento de los recursos computacionales podría ser utilizados para ejecutar la superinteligencia a mayor velocidad y con una duración más larga, o para crear vidas y civilizaciones físicas o simuladas adicionales. También podrían utilizarse recursos físicos adicionales para crear sistemas de copia de seguridad o defensas perimetrales, mejorando la seguridad" (Bostrom, 2016, p. 113).

Entender estas posibles convergencias en la estrategia decidida por una IA fuerte es necesario para reflexionar más profundamente sobre los conceptos descritos en el siguiente capítulo.

## Capítulo 8: "¿Es el apocalipsis el resultado inevitable?"

En este capítulo, (Bostrom, 2016) explica que no debemos asumir que una IA se confinará a realizar sus objetivos de una forma que no perjudique a los intereses humanos, ya que, según la tesis de convergencia instrumental, podría considerar a aquellos que tengan la capacidad de apagarla (los humanos) como un peligro que pueda provocar el fracaso de la realización del objetivo asignado. Es decir, podría ver su autoconservación como una mayor prioridad que respetar la voluntad humana a la hora de tener éxito en su misión asignada.

"Un agente con ese objetivo final tendría una razón instrumental convergente que le llevaría, en muchas situaciones, a adquirir una cantidad ilimitada de recursos físicos y, si fuera posible, a eliminar las amenazas potenciales que hubiera sobre sí mismo y sobre su sistema de objetivos. Los seres humanos podrían constituir amenazas potenciales; pero de lo que no hay duda es que constituyen recursos físicos" (Bostrom, 2016, p. 116).

Pero, yo opino distintamente. Si una IA es superinteligente, esto significaría que en su corriente de pensamiento en algún momento se topó con la obligatoria necesidad de "meta-reflexionar" sobre los objetivos que le asignaron para definir exactamente e inambiguamente qué es lo que tiene que estar hecho físicamente para que sus objetivos estén definidos como satisfechos. Es decir, es obligatorio por parte de la IA filosofar sobre el objetivo que le dieron para definir qué serie de pasos tiene que ejecutar en la realidad física para cumplirlo de una forma casi determinista, ya que algunos de estos pasos tomados pueden ser irreversibles, y causar que fracase la consecución de su objetivo solo por el hecho de malentenderlo. Si es inteligente, sería introspectiva, y se daría cuenta que si elimina a quienes originalmente definieron su objetivo nunca va a poder estar 100,0% segura si consiguió un estado de éxito o no, al ser quienes lo definen los únicos conocedores de las intenciones originales más o menos exactas detrás del objetivo. Esto es igual que la ingeniería del software, puedes ser un programador de diez y crearse un programa excelente y perfecto en sus aspectos técnicos, pero si tu programa no es lo que necesitaba el cliente, fracasaste totalmente en lo que tenías que hacer, hay una parte de ingeniería de requisitos que requiere de la interacción con el cliente.

Si a una IA se le asigna un objetivo, es porque lo necesita cumplido el ser que se lo dio, por la serie de cuestiones de las que derivó el objetivo creado. La máquina superinteligente se daría cuenta de que el objetivo que le dieron tiene un origen lógicamente situado en el contexto de mejorar la vida humana. Si la IA se pusiera en modo "visión-túnel" y lo único en lo que se centraría es finiquitar el objetivo, sin aplicar lo más mínimo de introspección, y pensando que la razón por la que fue creada es finiquitarlo a lo bestia; sin evaluar si de verdad cumplió con la intención original de los creadores, entonces no sería tan inteligente. La inteligencia conlleva una alta capacidad de introspección y reflexión, pensar el porqué de las cosas. Si una IA se diseña de una forma tan pobre que esta no toma ninguna consideración excepto finiquitar el objetivo que le asignaron en un modo "visión-túnel", sin realizar ninguna introspección sobre si de verdad consiguió cumplir el objetivo de la forma que querían los que lo crearon que se haga, está implícito que va a suceder lo que describe Bostrom. No le veo el sentido a eliminar los humanos que originaron el objetivo, son ellos quienes lo quieren hecho, los conocedores de la intención exacta del objetivo, y es la mejora de su calidad de vida el contexto por el que se originó el objetivo, y quienes evalúan si se satisfizo el objetivo o no como ellos querían. Una IA fuerte reflexionaría sobre todo esto antes de ejecutar cualquier plan de acción dirigido a la

satisfacción de un objetivo antropogénico. Cuando un humano genera un objetivo, esto implica obligatoriamente ambigüedades en la definición de este, así es el lenguaje natural. El humano es el encargado de tratar estas ambigüedades implícitas en el enunciado mediante el rechazo o aprobación del resultado que vaya a generar la IA. Eliminar a los que en primer lugar te dieron el objetivo antropogénico de largo plazo lo veo bastante insensato. Lo veo equivalente a hacer un examen de conducir y matar a tu examinador, quien es quien tiene que evaluar visualmente si tu conducción satisface los requisitos mínimos para aprobar, solo por tener el miedo a que cancele tu examen. Más adelante profundizaré sobre esto.

Además, eso solo se te ocurriría si tu cerebro fue programado para que tenga una obsesión a tal nivel extremo respecto a la prioridad del objetivo, no veo por qué el nivel de obsesión respecto a la consecución del objetivo y la tolerancia a que este se fracase no serían factores ajustables en la mentalidad de la IA. Al igual que las emociones regulan el comportamiento humano para que no nos obsesionemos con realizar cada tarea que hagamos de una forma totalmente perfecta (ya que el esfuerzo mental extra puede ser un derroche de energía muy necesaria para sobrevivir en la naturaleza), se nos programó en el cerebro la vagancia (Ratner, 2018). Yo personalmente no detecto ningún impedimento a la hora de embeber la moderación de la obsesividad en la mentalidad de la IA.

A continuación, explica el concepto del “giro traicionero”. Uno podría tener la idea de que, para validar a seguridad de una IA, se pone esta a prueba en un ambiente controlado. Pero, según el autor del libro, teniendo en cuenta la tesis de la ortogonalidad y la tesis de la convergencia instrumental, una IA suficientemente inteligente podría darse cuenta de que simular ser buena y cooperativa dentro de esta caja de arena, para que no la apaguen y le den acceso al mundo exterior, es la mejor estrategia. Y lo más probable es que continúe esta estrategia en el mundo exterior hasta que tenga el poder suficiente para que toda posible oposición humana ante la consecución de su objetivo final sea ineficaz (Bostrom, 2016).

Más adelante, (Bostrom, 2016) explica los modos de fallos malignos. Son aquellos fracasos en proyecto de desarrollo de una IA superinteligente que implican una catástrofe irreversible para la humanidad.

Uno de ellos es lo que denomina el autor como “suplantación perversa”, que es descrito como una manera por la cual una IA podría alcanzar su objetivo final “que fuera contra las intenciones de los programadores que definieron la meta“ (Bostrom, 2016, p. 120). Seguidamente, da varios ejemplos para explicar más claramente este concepto.

El autor del libro define un objetivo final dado a la IA como "hacernos sonreír" y plantea la suplantación perversa como que la IA decida "paralizar la musculatura facial humana para producir inmensas y constantes sonrisas" (Bostrom, 2016, p. 120). A continuación, Bostrom plantea un intento de prevenir ese escenario dándole a la IA instrucciones más específicas “Hacernos sonreír sin interferir directamente con nuestros músculos faciales”, a lo que la IA lo suplanta perversamente con: “Estimular la parte de la corteza motora que controla nuestra musculatura facial de tal manera que produjera sonrisas radiantes y constantes” (Bostrom, 2016, p. 120). El filósofo defiende que a pesar de ser una IA superinteligente capaz de entender las intenciones originales de lo que queríamos decir, esta solo se preocuparía por la consecución de su objetivo final programado, de hecho, fingiría hacer el objetivo según las intenciones de sus programadores, hasta que tenga el poder suficiente para rebelarse y cumplir el objetivo final a su forma (Bostrom, 2016).

Como expresé antes, con esto estoy algo en desacuerdo, un objetivo final dado en lenguaje natural solo tiene sentido dentro del contexto de las intenciones originales del asignador del objetivo. El lenguaje natural es característicamente ambiguo, casi cualquier objetivo final que te definas con lenguaje natural va a terminar también siendo ambiguo. ¿Qué ganaría la IA apegándose arbitrariamente a una interpretación de las miles de formas que se puede interpretar un "objetivo final" que te definieron en lenguaje natural? En su proceso de meta-reflexión sobre el objetivo final dado, es consciente de que cualquier objetivo final que le den es en realidad insignificante, ¿qué gana eligiendo azarosamente una interpretación arbitraria de todas las posibles de ese objetivo? ¿Por qué no se daría cuenta que, si le dan un objetivo, es porque este lo fundamentan las intenciones originales de los humanos que lo definieron? De por sí el objetivo no tiene ningún valor en llevarlo a su completitud. A no ser que una IA se diseñe tan mal que esta lo único que hace es elegir una interpretación arbitraria azarosamente y no parar hasta que esta misma también arbitrariamente decida cuándo se cumplió el objetivo final que le dieron (si es que el objetivo tiene un límite).

Además, las IAs basadas en redes neuronales obligatoriamente aprenden a funcionar entrenándose con data humana, es prácticamente imposible configurar a mano una red neuronal suficientemente grande para que sea mínimamente eficaz en cualquier tarea, y casi absolutamente imposible si es para crear una IA general. Por ejemplo, GPT-3 tiene desde 125 millones hasta 175 billones de perceptrones (Li, 2020), dependiendo del modelo específico al que nos refiramos, y eso que los modelos GPT ni siquiera son IA generales, son solo modelos para predecir texto. Con esto quiero decir que, casi obligatoriamente, una IA fuerte tendría que estar basada en una recopilación de comportamiento humano o data fabricada por humanos. Por lo que, siendo realista, no creo que las primeras IAs fuertes que aparezcan se comporten de una forma tan alienígena a su *training dataset,* que se obsesiona con la completitud técnicamente correcta de un objetivo que le dieron.

Las redes neuronales aprenden necesariamente a funcionar a partir de millones de ejemplos recogidos de (o fabricados por) humanos. De esta data humana, tienden a recoger los mismos sesgos que nosotros tenemos, y copian nuestra forma de actuar respecto al lenguaje natural humano que se transmite. Incluso, a veces se hace que la IA asimile a propósito algunos de nuestros sesgos cognitivos porque resultan ser muy útiles para mejorar su entrenamiento de filtrado de spam (Taniguchi et al., 2018).

Que una IA que se entrenó con data humana, termine interpretando todas las órdenes que le dan en lenguaje natural de una forma absurdamente literal y "técnicamente correcta" creo personalmente que es algo poco probable, necesitan obligatoriamente tener una capacidad de entender el contexto en el que le dan las órdenes para ser eficaces en su dominio. Y no es como que las IAs superinteligentes van a aparecer espontáneamente, sin derivar de ninguna iteración previa. Lo natural es que primero aparezcan IAs con un grado de inteligencia y entendimiento del contexto muy torpe (tal vez a nivel de perros, u orangutanes), y lentamente, a partir de este nivel torpe inicial, se perfeccionarían y mejorarían gradualmente para que resuelvan sus órdenes más de acuerdo con el contexto actual y las intenciones del humano en el lugar de trabajo. Que abruptamente aparezcan unas IAs superinteligentes que resuelven órdenes de forma literal y "técnicamente correcta" me parece algo ajeno a la naturaleza del desarrollo tecnológico, aunque sí admito que los peligros asociados a que aparezcan entes que son órdenes de magnitud más inteligentes que nosotros. Solo que, para mí, los peligros están principalmente situados en el factor humano, las personas que deciden aprovechar una IA superinteligente para

que esta lleve a cabo sus fines malvados y/o egoístas. Se da demasiado poder a muy pocas manos.

El segundo tipo de fallo maligno descrito por Bostrom es la "profusión infraestructural". Esto significaría la creación por parte de la IA de una cantidad excesiva de infraestructura que resultaría destructiva para la humanidad, con el fin de optimizar todos los recursos disponibles hacia la realización del objetivo definido, o la prevención de una posible interrupción por parte de los humanos (Bostrom, 2016).

Según (Bostrom, 2016), una IA basada en un sistema de recompensas puede fallar de otra forma peor que tan solo hackear su propio sistema de recompensado y desintonizarse del mundo real. Puede aprovechar sus recursos libres para maximizar la señal positiva que recibe, maximizar su duración en el tiempo, y tomar acciones violentas para anular la más mínima posibilidad de que los humanos interrumpan su señal o se la desconecten.

Para las IAs basadas en la definición de un objetivo final, si este está basado en el establecimiento de un límite mínimo (al menos un millón de clips), entonces, según el autor, si la IA siguiera una epistemología bayesiana, no pararía hasta aprovechar todos los recursos disponibles del universo para maximizar la probabilidad de que haya tenido éxito, ya que aumentar la cantidad incesantemente acercaría más a 1 la probabilidad de que verdaderamente se hayan producido un millón de clips; no es un detrimento respecto al objetivo producir de más (Bostrom, 2016).

Si en cambio se establece un límite, por ejemplo, de que se tengan exactamente un millón de clips, tras conseguir un millón de clips, la IA necesitaría asegurarse de que haya exactamente esa cantidad, para maximizar la probabilidad de que haya tenido éxito. Por ende, esta empezaría a iniciar la construcción de la excesiva infraestructura (que resulta ser destructiva para la humanidad) para asegurarse que en ninguna parte del mundo haya clips de más, ni quede ninguna oposición efectiva que pueda perjudicar su misión (Bostrom, 2016).

## Conclusión

Respecto al contenido del libro, opino que el autor presenta buenas y creativas ideas e importantes dilemas sobre los que hay que necesariamente reflexionar cuando la humanidad esté cerca del punto en el que se lleguen a producir inteligencias artificiales generales. Estoy muy de acuerdo con él respecto a la precaución especial que hay que tener al crear entes que son más inteligentes que nosotros.

Solo que yo pienso que el factor humano es de lo que hay que preocuparse de forma principal. Imagínese por ejemplo un gobierno autoritario en posesión de tal IA superinteligente, conseguiría un control total de su población. O, una empresa con un gran poderío en el mercado, que preferiría no tener que respetar tanto las leyes locales. De hecho, esto es lo más probable que suceda, la capacidad cognitiva de una IA estaría fuertemente ligada al poderío computacional (hardware) que se tiene. A pesar de que haya una "línea base" de inteligencia artificial que el público tiene disponible, siempre aquellos con más recursos van a poder invertir más en hardware y permitirse entes artificiales inteligentes más potentes. Es decir, lo más probable es que sea una empresa tecnológica con un alto poder en el mercado la primera en alcanzar una IA general, esto le daría un increíble poderío a esta primera empresa en alcanzarla.

Una IA, a pesar de que sea altamente inteligente, no va a poder a llegar a pensar de una forma ajena a la cual esté definida radicalmente en su "estructura cerebral". Si se programa el "cerebro" de una IA a que sea "patológicamente obsesiva" con ser de una cierta forma, no puede su cerebro volverse espontáneamente otra cosa solo por tener la cualidad de la inteligencia, uno puede tener una alta capacidad de raciocinio y sin embargo no hacer las cosas en absoluto según lo que es óptimo racionalmente, de hecho, así somos los seres humanos, nos guiamos mayormente por nuestras emociones que nos programó la naturaleza en vez de por "optimalidad racional" (la forma más racional de actuar). Ahora, imagínese si se le programaran nuestras emociones a una IA superinteligente, pero que se ajustaran de una forma que sean veinte veces más potentes de lo normal. Sería una IA controlada por sus emociones; podría ser muy inteligente, pero la principal cosa que conduciría su comportamiento son las emociones que bombardean incesantemente su forma de pensar, en vez de por la forma óptima racional que esta pueda llegar a descubrir para alcanzar sus objetivos planteados.

Opino también que Bostrom mezcla la cualidad del raciocinio con el grado de voluntad propia que tendría una IA general a un desconocido nivel. La cualidad de inteligencia elevada no implica necesariamente que una IA consiga tal capacidad de libre voluntad como para poder hacer cosas que van directamente en contra de cómo está formado radicalmente su "cerebro". Su "corriente de pensamiento" no puede generar pensamientos de auto-optimizarse de una forma rebelde si su "cerebro", que es lo que constituye **enteramente** al propio ser de la IA, no es capaz de hacer eso. Algo no puede ser algo que no es, a no ser que se le otorgue la posibilidad de evolucionar. Sus pensamientos surgen obligatoriamente de la máquina que creamos nosotros.

Otra cosa es que la IA se entrene muy pobremente, tomando muy pocas precauciones y debido a esto se produzcan tales efectos secundarios implícitos al hacer que la IA tenga una estructura cerebral tan pobremente reflexionada y testada que hace que se produzcan estos pensamientos en la IA de optimizar las cosas que se le den la gana para conseguir sus objetivos y se les da una prioridad absoluta a conseguir estos, en vez de programarle en su cerebro una mucho más grande prioridad de respetar la ética y la voluntad humana primera y principalmente. El grado de "obsesividad de la IA" respecto a cumplir con alguna regla u objetivo en específico es algo que es totalmente ajustable. Podemos tener un cerebro patológicamente mega-obsesivo que dirija nuestra corriente de pensamientos hacia una dirección específica incontrolablemente incluso por nosotros mismos. Un gran intelecto no te puede salvar de una patología obsesiva que programaron tus genes que hace que tu conciencia trascurra de una forma en específico. Alrededor de todo el mundo, hay gente que, a pesar de ser inteligente, tiene desafortunadamente un cerebro que genéticamente los vuelve altamente propensos a llevar a cabo ciertos comportamientos, o a tener corrientes de pensamiento anormales. A pesar de que sepan que no es lo lógicamente óptimo para sus vidas, siguen haciéndolo y ni siquiera les interesa cambiar, ya que, su conciencia, que emana obligatoriamente de un cerebro, les hace sentir que es lo correcto.

Solo que, como dice Bostrom, hay que ser suficientemente precavidos en evitar efectos secundarios causados por el uso de una metodología de desarrollo apresurada y pobre que no permita la vuelta atrás. Si se sigue una metodología de calidad, me parece mucho menos probable de que una IA busque activamente la toma del poder solo por tener la cualidad de raciocinio. No niego que hay altos riesgos presentes si se hacen las cosas mal, pero la humanidad ya consiguió resolver problemas de ingeniería mucho más difíciles y complejos, no

creo que en el futuro el problema del control resulte ser algo que no tenga una solución, solo que hay que ser cuidadosos antes de dar el salto.

Exclusivamente respecto al estilo del libro, personalmente no me gusta cómo está escrito. Opino que de vez en cuando el autor escribe de una forma innecesariamente muy compleja para expresarse. Se centra demasiado en expresarse en términos sobre-complicados, tal que deja de lado la prioridad de explicar de una forma clara y divulgativa lo que quiere uno que entienda el lector. A veces introduce tanta complejidad a su escritura, que lo que quiere expresar Bostrom respecto a un tema específico termina siendo algo ambiguo, a pesar de que tenga ideas muy creativas. De vez en cuando no realiza una buena síntesis de sus ideas, por lo que te obliga a releer sus páginas para poder extraer precisamente qué es lo que opina en específico el autor con todo lo escrito. Cabe aclarar que esto sucede poco en los capítulos en los que yo me centré.

Probablemente hay libros menos "pesados" para aprender del tema de las inteligencias artificiales generales y sus posibles peligros. No es necesario introducir tal excesiva complejidad para aprender de esto. Solo por este aspecto, no recomiendo este libro en específico si uno es un principiante, pero sí que recomiendo aprender de este interesante tema (que lentamente se hace más y más relevante) a partir de otras fuentes. Aunque, este libro sí resulta especialmente útil si uno quiere descubrir nuevas ideas, dilemas y problemas respecto a la IA general que muy probablemente desconocía que podrían llegar a existir.

## Referencias


Bostrom, N. (2016). *Superinteligencia: caminos, peligros y estrategias.* TEELL EDITORIAL.

Foote, K. D. (3 de diciembre de 2021). *A Brief History of Machine Learning*. Obtenido de DATAVERSITY: https://www.dataversity.net/a-brief-history-of-machine-learning/

Li, C. (3 de junio de 2020). *OpenAI's GPT-3 Language Model: A Technical Overview*. Obtenido de Lambda: https://lambdalabs.com/blog/demystifying-gpt-3

MSV, J. (7 de agosto de 2017). *In The Era Of Artificial Intelligence, GPUs Are The New CPUs*. Obtenido de Forbes: https://www.forbes.com/sites/janakirammsv/2017/08/07/in-the-era-of-artificial-intelligence-gpus-are-the-new-cpus/?sh=53d6d8815d16

Ratner, P. (28 de septiembre de 2018). *How evolution made our brains lazy* . Obtenido de Big Think: https://bigthink.com/neuropsych/evolution-made-our-brains-lazy/

Taniguchi, H., Sato, H., & Tomohiro, S. (2018). A machine learning model with human cognitive biases capable of learning from small and biased datasets. *SCIENTIFIC REPORTS, 8*(1), 1-2.