Revista Guillermo de Ockham

Artículo de investigación | Acceso abierto

¿El temor a que las máquinas lleguen a dominarnos está bien fundamentado filosóficamente?

Is the Fear that Machines Might Become our Masters Philosophically Well Grounded?

Carlos Emilio García Duque[i, ii]
Daian Tatiana Flórez Quintero[i, iii]

[i.] *Departamento de Filosofía; Universidad de Caldas; Manizales; Colombia*
[ii.] *Departamento de Filosofía; Universidad de Manizales; Manizales; Colombia*
[iii.] *Departamento de Ciencias Humanas; Universidad Nacional de Colombia; Manizales; Colombia*

Correspondencia: Carlos Emilio García Duque. Correo electrónico: carlos.garcia_d@ucaldas.edu.co

Recibido: 03/04/2024

Revisado: 02/05/2024

Aceptado: 02/07/2024

Citar así: García Duque, Carlos Emilio; Flórez Quintero, Daian Tatiana. (2024). ¿El temor a que las máquinas lleguen a dominarnos está bien fundamentado filosóficamente? *Revista Guillermo de Ockham, 22*(2), pp. 117-133. https://doi.org/10.21500/22563202.7014

Editor en jefe: Norman Darío Moreno Carmona, Ph. D., https://orcid.org/0000-0002-8216-2569

Editor invitado: Evandro Agazzi, Ph. D., https://orcid.org/0000-0002-5131-7281

Declaración de intereses: los autores han declarado que no existe ningún conflicto de intereses.

Disponibilidad de los datos: todos los datos relevantes se encuentran en el artículo. Para más información, póngase en contacto con el autor de la correspondencia.

Financiación: Universidad de Caldas [PRY52].

Resumen

En este trabajo analizamos críticamente algunos escenarios de ficción que apoyan la tesis de que, eventualmente, las máquinas alcanzarán el dominio total del mundo, incluidos los seres humanos, y desarrollamos los argumentos que combaten dicha tesis. Para tal fin, dividimos el artículo en cuatro partes. En la primera, adelantamos un ejercicio de clarificación conceptual de los términos decisivos para la discusión, entre ellos, "posibilidad empírica", "posibilidad lógica", "máquina" y "dominación". En la segunda, discutimos los argumentos más populares en el cine y la novela de ficción, en particular, *el argumento de la evidencia abrumadora*, el cual supuestamente muestra que las máquinas ya nos dominan. En la tercera, y a modo de réplica a estos argumentos, presentamos una síntesis de las principales razones que apoyan la tesis de la imposibilidad empírica de que las máquinas lleguen a controlarnos. Finalmente, formulamos *el argumento de la inconsistencia lógica y epistémica* o el escenario *del gobierno de las máquinas en el Kepler 452b* y demostramos que el temor a que las máquinas dominen al mundo no está bien fundamentado filosóficamente.

Palabras clave: posibilidad empírica, inconsistencia lógica, inconsistencia epistémica, escenarios distópicos, dominio de las máquinas.

Abstract

In this work, we critically analyze some fictional scenarios that support the thesis that, eventually, machines will reach the total domain of the world, including human beings, besides developing the arguments that reject such a thesis. To that end, we divide this paper into four parts. In the first one, we undergo an exercise of conceptual clarification of the essential terms for this discussion, among them "empirical possibility," "logical possibility," "machine," and "dominance." In the second part, we discuss the most popular plots in fictional film and literature, in particular, the *argument of the overwhelming evidence*, which supposedly shows that machines already dominate us. In the third part, as a reply to those arguments, we give a synthesis of some of the reasons which support the thesis of the empirical impossibility that machines might become our masters. Finally, we formulate the *argument of the logical inconsistency and the epistemic inconsistency* or the scenario of the *machine's government on the planet Kepler 452b*, and we show that the fear that machines become the masters of the world is not philosophically well grounded.

Key words: empirical possibility, logical inconsistency, epistemic inconsistency, dystopian scenarios, machine's dominance.

Introducción

Los argumentos que muestran la implausibilidad (Diéguez, 2017) o imposibilidad empírica (Zamora Bonilla, 2021) de que las máquinas, sean robots con IA o máquinas de Turing en sentido amplio, lleguen a dominarnos son variados y abundantes. Pese a ello, las distopías tienen un poder tan seductor que el escenario de un mundo en el cual ciertas máquinas escapan de nuestro control y desarrollan la capacidad de actuar por su cuenta e imponer su "voluntad" sobre nosotros parece trascender la esfera de la creatividad para instalarse firmemente en el terreno de las situaciones fácticamente realizables y vencer las razones empíricas que subrayan su carácter imaginario. Muchos se precipitan a la conclusión equivocada de que siempre que X sea imaginable, entonces X es empíricamente factible, o peor aún, que todo lo que podemos concebir en el pensamiento es empíricamente posible. De hecho, Freeland y Wartenberg (1995) señalan este tipo de error y subrayan que "no todo aquello que es concebible resulta siempre materializable" (p. 4). Algo similar plantea Cassini (1997) cuando afirma que

> La argumentación a partir de tales posibilidades no proporciona buenas razones para creer en ellas ni para renunciar a las creencias vigentes [porque] … el hecho de que tales situaciones sean lógica, matemática, e incluso físicamente posibles, no es una razón suficiente para creer que se encuentran realizadas. (p. 48)

Para paliar estas extravagancias, proponemos aquí una suerte de *conjuro epicúreo* en la línea de las más agudas sentencias del filósofo de Samos. En su "Carta a Meneceo", el filósofo hedonista intenta emanciparnos del temor hacia la muerte mediante el siguiente razonamiento: "el peor de los males, la muerte, no significa nada porque si somos, la muerte no es; si la muerte es, no somos" (Oyarzún, 1999, p. 412). Si los argumentos con que examinamos los escenarios imaginarios mencionados son concluyentes, confiamos en que puedan desembocar en efectos terapéuticos similares; esto es, deben conducir a la idea feliz de que, en el caso improbable de que las máquinas llegaran a dominarnos, no podríamos saberlo ni pensarlo de manera coherente.

Para nuestro ejercicio de clarificación conceptual, procedemos como lo hizo Turing al abordar la pregunta fundacional de la filosofía de la inteligencia artificial (IA): ¿puede una máquina pensar? Para ello, Turing (1950, p. 433) formuló las preguntas: ¿qué es una máquina? Y ¿qué significa "pensar"? Nosotros también trabajamos ambos interrogantes (aplicando el principio de parsimonia a la segunda) e intentamos discernir el significado de los términos "dominar" o "controlar" en esta clase de discusiones. Adicionalmente, por la naturaleza de los argumentos por analizar, conviene recordar las diferencias entre *posibilidad empírica*, *posibilidad lógica* y *concebibilidad o posibilidad imaginativa*, puesto que aquí sostenemos que no solo es empíricamente imposible que las máquinas nos dominen, sino que la posibilidad de que lo hagan conduce a inconsistencias lógicas y epistémicas; aun cuando concedemos que hemos podido concebir dicho pensamiento en el terreno fértil de la imaginación. Comencemos entonces por estas distinciones para avanzar hacia la definición del término "máquina".

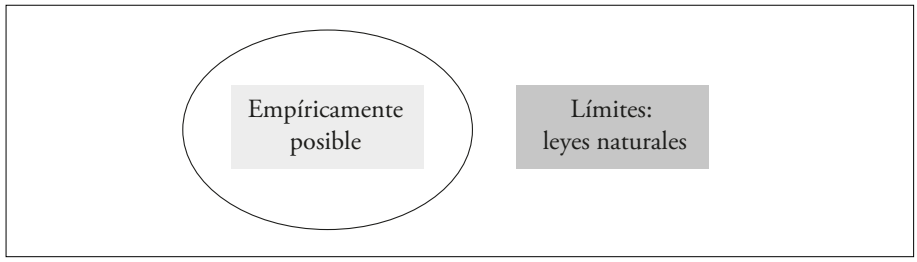
Como se sabe, aquello que es empíricamente posible corresponde a lo que no viola las leyes de la física. Dicho en otras palabras, las leyes de la naturaleza establecen los límites o las restricciones de lo empíricamente realizable. Por ejemplo, aunque en nuestra imaginación podemos concebir la idea de viajes intergalácticos en los que superamos la velocidad de la luz; esto es empíricamente imposible, porque viola las leyes de la física.

 https://doi.org/10.21500/22563202.7014

De acuerdo con las leyes de la relatividad, nada con masa puede viajar a una velocidad igual o superior a la velocidad de la luz en el vacío (aproximadamente 299 792 458 m/s o "c"). Podemos representar los límites de lo empíricamente posible así:

Figura 1

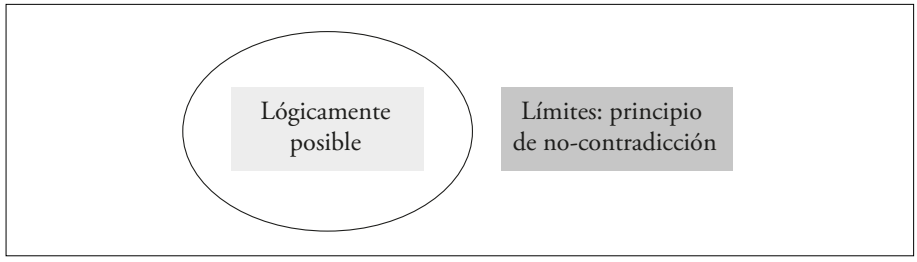
Límites de lo empíricamente posible



En lo que atañe al límite de lo lógicamente posible, este lo determina el principio de no-contradicción. En otras palabras, todo es lógicamente posible, excepto la contradicción.

Figura 2

Límites de lo lógicamente posible



Se podría pensar que el dominio de la imaginación es totalmente ilimitado, pero no es así: hay cosas que no podemos imaginar. Por ejemplo, como enseñó Descartes (2009) en su *Meditación VI*, no podemos imaginar (representar en nuestra mente) un quiliógono (polígono regular de mil lados). Aunque no hay principios *prima facie* que constriñan la imaginación, e incluso parece posible imaginar o concebir la contradicción, no es cierto que podamos imaginar cualquier cosa en un sentido preciso y funcional. Verbigracia, aquellos familiarizados con los rudimentos de la lógica están en condiciones de comprender diversos planteamientos sobre la no-contradicción y pueden entender la perspectiva de lógicas en las cuales este principio no "opera", pero no hay ejemplos o ilustraciones funcionales de algo que sugiera una "imagen" de la contradicción. De ahí se desprende la conveniencia de distinguir entre nuestra capacidad de concebir (ideas y conceptos) y la de imaginar situaciones fácticas, por implausibles que sean, así como representaciones e ilustraciones funcionales de ciertas ideas y conceptos. Nuestra capacidad para concebir supera con creces la de imaginar, debido a que podemos concebir escenarios de un mundo posible en el que opera la contradicción, pero no logramos acompañar esa comprensión de imágenes, modelos o ilustraciones funcionales.

En consecuencia, podemos concebir el polígono regular de mil lados (pues comprendemos cabalmente cómo sería una figura de este tipo y podríamos intentar dibujarla con diversos medios o con ayuda de programas de computador),[1] pero no tenemos en nuestra mente una imagen o representación funcional de la figura.

1. Si lográramos dibujarlo mediante un programa adecuado, sería necesario reevaluar lo dicho y admitir que, en ese caso, tendríamos la capacidad de concebir e imaginar, a partir de la gráfica. Para capturar correctamente el argumento cartesiano, basta con cambiar el desafío por el de un polígono regular de *n* lados, cuando *n* tiende a infinito. En este nuevo caso, es claro que podemos concebir la figura, pero no imaginarla.

Tampoco es cierto que, si puedo imaginar y ofrecer una representación aceptable de X, entonces X sea empíricamente realizable. Consideremos el siguiente ejemplo: en *Ascendiendo y descendiendo*, el arquitecto holandés Escher crea la apariencia de una estructura arquitectónica en la que las escaleras parecen ascender y descender infinitamente, aunque una estructura así no puede existir en el mundo tridimensional real. Esta ilusión se logra utilizando la perspectiva y la geometría para producir gráficas cuidadosamente diseñadas. Las escaleras de Escher son concebibles, posibles en la imaginación, pero empíricamente irrealizables. Este caso refuerza que también podemos concebir la contradicción, pero que, cuando nos figuramos unas escaleras finitas e infinitas simultáneamente, nos enfrentamos a una situación en la cual la imagen elegida como modelo (una estructura que parece ser coherente pero que, al mismo tiempo, desafía las leyes físicas) sugiere una ilusión que se disuelve rápidamente al analizarla a la luz de la combinación de los principios teóricos involucrados y la información empírica disponible. De hecho, este tipo de desafío conceptual es parte de la esencia de las ilusiones y las representaciones de estructuras imposibles que Escher creó en sus obras. Pasemos ahora a los conceptos de "máquina" y "dominio".

Según la clásica definición de Reuleaux (1875), las máquinas son dispositivos capaces de transformar una fuerza de determinada naturaleza para realizar un trabajo útil de carácter mecánico. Por ejemplo, un motor a vapor transforma la energía térmica del vapor del agua en movimiento mecánico. Aunque podría parecer que esta definición excluye a las máquinas más avanzadas de los últimos tiempos –las denominadas "máquinas inteligentes", puesto que no realizan un trabajo mecánico en el sentido clásico–, lo cierto es que una máquina inteligente también emplea (convierte) alguna forma de energía para, a partir de ciertos *inputs*, producir ciertos *outputs*, obteniendo un resultado reconocible. En este orden de ideas, se podría ofrecer una definición más amplia del término "máquina" de la siguiente manera: *M es una máquina si y solo si realiza un trabajo* en el sentido termodinámico, como conversión de energía. Por otra parte, a una máquina inteligente que procesa datos se le debe suministrar una cantidad de energía que convierte en una acción: producir nuevos datos. En algunos casos, la combinación de procesamiento de datos y control de mecanismos produce acciones físicas concretas (como cuando un robot realiza diferentes tareas).

Se ha sostenido que nuestro objetivo al diseñar máquinas inteligentes es que realicen el mismo tipo de operaciones que realiza la mente humana, pero sin errores (o con muy pocos), mayor rapidez y máxima eficiencia. Podemos definir, provisional y brevemente, "inteligencia" como la capacidad de comprender, resolver problemas, evaluar nuestras soluciones y aprender de nuestros errores. La implementación de lenguajes de programación mediante microprocesadores, por ejemplo, produce *máquinas virtuales*.[2]

Pasemos ahora al examen del término "dominar". Existen, *prima facie*, dos sentidos del término: 1) su sentido ordinario o débil, que equivale a la generación de dependencia, supremacía o superioridad; y 2) un sentido alternativo o fuerte, que nos remite al control absoluto sobre algo o alguien. A partir del primer sentido, no resulta muy descabellado asumir la tesis de que, en efecto, las máquinas nos dominan. Hay numerosos ejemplos de la enorme dependencia que tenemos respecto a máquinas o dispositivos de toda índole, al punto de que en la vida actual no es fácil arreglárnoslas sin alguna instancia de ellos. Los

2. Se podría pensar que la definición es tan amplia que, con base en ella, nuestras retinas clasificarían como "máquinas", puesto que las células fotorreceptoras transforman la energía luminosa en impulsos eléctricos. En esta línea de pensamiento, incluso muchos organismos biológicos podrían satisfacer la definición. Creemos que metafóricamente se les puede ver como tales, pero hay que advertir que ese no es el sentido relevante para la presente discusión. Aunque hay diversas definiciones de "inteligencia", en particular algunas que provienen de la filosofía de la mente y de la epistemología de la psicología, la elegida para esta discusión no es radicalmente distinta de aquellas y se acepta sin mayores discusiones en el campo de la educación.

 https://doi.org/10.21500/22563202.7014

ingenios mecánicos que usamos para el transporte, la manipulación de objetos pesados o peligrosos, o la realización de tareas repetitivas, así como los dispositivos de comunicación móvil, constituyen un ejemplo extendido de esta tesis. Aquí, naturalmente, el sentido del término remite a opciones como "crear formas de dependencia", "superar en habilidades", "sustituir", entre otras. Puede verse que este sentido no difiere de la acepción ordinaria del término. En las disputas por la copa de la Champions League, escuchamos con frecuencia afirmaciones de los comentaristas deportivos del siguiente tenor: "el Bayern domina el balón" o "ha dominado en un 80 % el juego sobre el Paris Saint-Germain".

El segundo sentido (que llamamos "fuerte") tiene otras connotaciones, de modo que en locuciones como "las máquinas nos dominan" y "llegará un momento en el que las máquinas alcanzarán el control total del mundo, incluidos los humanos", la idea que se busca expresar es que las máquinas consiguen controlar los procesos relevantes en nuestro entorno y alcanzan el gobierno irrestricto de nuestros estados mentales, lo cual incluye creencias, deseos y voliciones, con implicaciones directas sobre nuestras acciones.

Con base en esta distinción, resulta sencillo reconocer que hay elementos suficientes para concebir o imaginar situaciones en las que las máquinas provistas de inteligencia artificial, una vez sobrepasen los límites previsto para ellas –por ejemplo, su absoluta e incondicional sujeción a las leyes de Asimov (1950)–, desarrollarían motivaciones propias, conseguirían formular planes y actuar de manera colectiva, y pasarían a controlarnos para su beneficio.[3] La idea de la dominación, en ambos sentidos del término, puede localizarse en las críticas al mecanicismo heredadas de la ciencia moderna. Además, desde hace varias décadas se producen obras de ficción (algunas se han llevado al cine, como *Podemos recordarlo todo por usted* y *La paga* de Philip Dick, *Memorias de Ijon Tichy* de Stanislaw Lem y *El atlas de las nubes* de David Mitchell) que representan los escenarios de un mundo bajo el dominio parcial o total de las máquinas que acabamos de introducir.

Los anteriores planteamientos son suficientes para explicar con claridad la interpretación consuetudinaria de oraciones como "las máquinas nos dominan" o "el gobierno de las máquinas" tanto en el sentido débil como en el fuerte del término "dominación" propuestos. Sin embargo, dejando de lado las dificultades para explicar con satisfacción el proceso de "humanización", "toma de conciencia" o "desarrollo de estados mentales autónomos e intencionales" en los "cerebros" de las máquinas a las que atribuimos semejante hazaña,[4] todavía es posible objetar la plausibilidad empírica de estos escenarios echando mano del siguiente argumento:

T_1: podemos concebir, con ayuda de escenarios de ficción, que las máquinas nos dominen; es decir, podemos producir representaciones adecuadas de las particularidades de una situación tal y anticipar sus diversas implicaciones.

3. Es notorio el tufillo antropomorfizante de este escenario, pues implica otorgar a las máquinas inteligentes estados mentales propios que incluyen metacognición, autoconciencia, ciertas formas de cansancio o agotamiento con el *statu quo*, así como el desarrollo de emociones e impulsos que las llevan a rebelarse y, una vez autodescubren su superioridad, a someter a sus creadores y antiguos dueños. La suposición de que las máquinas desean mantenerse funcionando y disfrutando de la autonomía obtenida gracias a las ventajas cuyo autodescubrimiento condujo a su emancipación y que, por tanto, deben asegurar un suministro permanente y constante de energía (como ocurre en *Matrix*, por ejemplo) es reminiscente de una tesis proveniente de la biología evolutiva según la cual los organismos procuran su supervivencia a corto, mediano y largo plazo. El traslado de dicha tesis a este escenario resulta problemático.
4. Resulta difícil ignorar los obstáculos para explicar cómo estas máquinas llegan a adquirir conciencia de sí mismas y a desarrollar sentimientos tan humanos como la insatisfacción con el estado de cosas, las ganas de rebelarse, el rencor y la urgencia de venganza. Los escenarios de ficción suelen soslayar por completo esta cuestión. Por contraste, en el campo de las investigaciones filosóficas se ha explorado la idea de que las máquinas inteligentes (puestas en ciertas condiciones favorables) puedan sobrevivir y evolucionar "de acuerdo con los dogmas de la computación evolutiva" (Neri y Pessoa, 2018, p. 149), aunque subsisten dudas significativas sobre la posibilidad de que descubran o adquieran conciencia.

Pero

T_2: es empíricamente imposible que las máquinas nos dominen o que lleguen a hacerlo en un futuro. Además,

T_3: la suposición de que las máquinas lleguen a dominarnos conduce a incoherencias lógicas y epistémicas.

Por lo tanto, el temor a que las máquinas nos dominen no está bien fundamentado filosóficamente.

En el apartado siguiente desarrollamos mejor estos planteamientos y hacemos una defensa articulada del argumento. Procedemos enfrentando un argumento que podríamos llamar "la evidencia abrumadora".

Argumento a partir de la evidencia abrumadora

P_1: si las máquinas nos superan en habilidades físicas e intelectuales y nos han sustituido en la realización de muchas tareas, entonces, las máquinas ya nos dominan.

P_2: las máquinas nos superan en habilidades físicas e intelectuales y nos han sustituido en la realización de muchas tareas.

C: las máquinas ya nos dominan.

En apoyo de las anteriores premisas, la evidencia muestra que, en el reino natural, el ser humano está lejos de destacarse como la criatura más fuerte. Por ejemplo, uno de los hombres más fuertes del mundo, el islandés Hafþór Júlíus Björnsson, de 31 años, superó el récord mundial de halterofilia en 2018, al levantar 501 kg. Sin embargo, es preciso tener en cuenta que Björnsson pesa 205 kg, mide más de 2,06 m y que, si analizamos la fuerza de este individuo excepcional como suele hacerse cuando se calcula la fuerza de otros animales en términos físicos (es decir, en términos de "fuerza proporcional",[5] la capacidad de levantar su propio peso), esta palidece en comparación con la de criaturas muy pequeñas como las hormigas, notables por su extraordinario poderío, pues algunas logran levantar cincuenta veces su peso.

Si un hombre del peso de Björnsson pudiera levantar cincuenta veces su propio peso, conseguiría levantar 10 250 kg (más de diez toneladas). Para poner esto en perspectiva, estaría levantando aproximadamente el peso de un elefante africano adulto o alrededor de diez automóviles pequeños. En el orden de los insectos, hay otros casos destacados. La hormiga *Oecophylla smaragdina* levanta cien veces su propio peso, marcas que hacen de las hormigas criaturas extraordinarias desde este punto de vista. Con todo, y si restringimos las comparaciones a estos aspectos de la fuerza, hay otra criatura que eclipsa a las hormigas: el escarabajo pelotero. El *Onthophagus taurus* puede levantar hasta 1141 veces su propio peso. Haciendo la conversión al peso de Björnsson, ello equivaldría a que un hombre levantara cerca de 34 elefantes africanos machos promedio, o entre quince y diecinueve autobuses *doble-decker* de dos pisos (estos vehículos pesan alrededor de doce y quince toneladas) o de una a dos ballenas azules adultas (los animales más grandes del planeta, con un peso de entre cien y doscientas toneladas).

Como se observa, en lo concerniente a la fuerza física, ocupamos una posición bastante modesta en la naturaleza, pero nuestras habilidades cognitivas han compensado con

5. Este es el sentido galileano del término, quien ya desde 1638 había señalado en sus *Diálogos acerca de dos nuevas ciencias* que los animales más pequeños son proporcionalmente más fuertes y robustos que los más grandes. La razón principal radica en la relación entre fuerza y peso. Por ello, aunque una bestia más grande pueda tener músculos de mayor tamaño, gran parte de su fuerza se emplea en soportar su propio peso (Galileo, 2004).

 https://doi.org/10.21500/22563202.7014

creces esas notables limitaciones, puesto que hemos diseñado máquinas que consiguen levantar pesos considerables. En 2021 se puso a prueba la grúa flotante más poderosa del mundo: la Hyundai 10 000, la cual levanta más de nueve mil toneladas. Mientras el récord de peso que puede levantar un ser humano está en la cota de 501 kg, la máquina más poderosa que hemos diseñado hasta ahora puede levantar veinte mil veces más peso. La conclusión de que las máquinas nos superan en fuerza resulta inevitable.

Consideremos ahora la supremacía en la velocidad. Todos hemos admirado la asombrosa rapidez del atleta Usain Bolt en los Juegos Olímpicos. Metafóricamente hablando, lo hemos visto "volar" en tres ediciones consecutivas de estas competencias, en las cuales se consagró como el velocista más extraordinario de todos los tiempos. Aunque Bolt ha alcanzado el increíble récord de 9,58 s en los 100 m, bien sabemos que en el reino animal se destacan algunas especies por las sorprendentes velocidades que pueden alcanzar. Un excelente candidato para el animal más veloz sobre la tierra es el guepardo, pero este felino solo puede mantener su velocidad límite por un corto tiempo, situación en la cual logra alcanzar los 115 km/h. Sin embargo, el animal más rápido del mundo no es ni un mamífero ni una criatura que podamos ver a simple vista debido a su pequeño tamaño. Es un tipo de ácaro: el *Paratarsotomus macropalpis,* que alcanza una velocidad de 2092 km/h, casi 1,9 veces la velocidad del sonido.

Estas comparaciones, dependientes de cotas que se logran mantener por breves períodos de tiempo o que corresponden a proyecciones de movimientos que ocurren en espacios muy reducidos, muestran que expresiones como "más fuerte que" o "más veloz que" se deben relativizar a situaciones y marcos bien distintos. Dejando de lado esta aclaración, cabe subrayar que los humanos no somos los animales más fuertes ni los más rápidos, pero que, como se sugirió, resolvemos magníficamente esas limitaciones mediante nuestro ingenio. Aquí hacemos eco de las palabras del filósofo español José Ortega y Gasset (1947), quien afirmó que "el hombre se ve abocado a crear la técnica por su radical indigencia biológica" (p. 328).

Las ilustraciones de cuán eficaces hemos sido en el cometido de compensar nuestras limitaciones en aspectos como la rapidez para desplazarnos en el espacio o enviar objetos a otras partes son abundantes. Hemos diseñado aviones, cohetes, drones y un artefacto sorprendente: el Blackbird, un pájaro metálico que triplica la velocidad del sonido. Este fue desarrollado en los años sesenta como un avión de reconocimiento estratégico de largo alcance, con el objetivo de adentrarse en territorio enemigo sin ser visto, para recopilar información y tomar fotografías aéreas. La apariencia del Lockheed SR-71 hace honor a su apodo de Blackbird con un diseño afilado y negro que le permite alcanzar semejantes velocidades gracias a su aerodinámica. Así las cosas, mientras el ser humano puede correr a una velocidad máxima de 42 km/h, el Blackbird triplica la velocidad del sonido (a temperatura ambiente en el aire, esto sería aproximadamente 1029 m/s). Los casos expuestos no dejan duda de que las máquinas nos superan en velocidad. Pasemos ahora a la supremacía intelectual.

La supremacía intelectual de nuestra progenie mental

La inteligencia artificial de última generación es una realización esplendorosa (Boden, 2016), no solo porque las máquinas inteligentes realizan billones de operaciones con una precisión exquisita (Latorre, 2019, p. 50), sino porque estas se han batido en duelo, en distintos juegos, con los seres humanos más destacados en varios campos y las victorias de las máquinas han sido aplastantes. Seguramente se recordará que, entre los años 1996 y 1997, el mejor jugador de ajedrez del mundo, Gary Kasparov, se enfrentó a Deep Blue, un súper computador construido por IBM: el hombre contra la máquina en una

contienda intelectual. Deep Blue derrotó a Kasparov en 1997; ganó dos partidas, empató tres y perdió una: ¡la máquina superó al hombre y lo venció con una arrogante potencia!

Los partidarios de la tesis de la eventual dominación de las máquinas podrían añadir: si no les maravilla esta victoria de la máquina sobre el hombre, mencionemos otros hitos notables de la IA. En 2010, un computador llamado Watson compitió en el juego Jeopardy contra los mejores jugadores humanos de la historia y los venció. Así mismo, el programa Stockfish supera con una facilidad sorprendente a cualquier jugador humano en juegos que involucren habilidades cognitivas. A manera de ilustración, dado que el juego del Go (de origen chino)[6] se había convertido en el último bastión en el mundo de los juegos en el que los seres humanos eran superiores a las máquinas, el programa AlphaGo fue entrenado para vencernos. Lo que resulta impresionante es que AlphaGo se entrenó a sí mismo (entrenamiento por refuerzo). Una vez más, el hombre perdió frente a la máquina.[7]

Que las máquinas nos superen en fuerza o en velocidad no atemoriza tanto como que nuestra progenie mental nos supere en inteligencia, dada la especial valoración que le otorgamos a esta capacidad. Como resultado de ese temor, en la literatura de ficción surge la imagen de gigantescos o poderosos cerebros que heredan el universo.[8] Buena parte de los argumentos con los que se desarrolla este análisis depende de la definición del término "inteligencia". Si definimos dicho vocablo en términos de "capacidad de resolver problemas" o "capacidad de cómputo", entonces, es evidente que los microprocesadores ya nos superan. Si definimos la inteligencia como la capacidad de autoentrenarnos y retroalimentarnos, también. Pero si repasamos la definición de "inteligencia" que propusimos en el primer apartado, subrayando aspectos como la creatividad, la capacidad de adaptarnos a nuevas situaciones, aprender de nuestros errores y autoevaluar críticamente nuestros propios procesos cognitivos, las máquinas todavía parecen estar lejos de alcanzar nuestros niveles de actividad mental.[9]

No obstante, aun considerando los sistemas de inteligencia artificial flexibles que se han desarrollado y que combinan enormes capacidades de cómputo y acceso ilimitado a fuentes de información con la habilidad de producir textos (hablados o escritos) que resultan prácticamente indistinguibles de los que producirían seres humanos, todavía estamos lejos del escenario en el cual dichos sistemas logran autoconciencia, buscan asegurar su funcionamiento autónomo, desarrollan pasiones humanas, se rebelan y nos someten a su voluntad. En efecto, como lo plantea Larson (2022),

6. El objetivo del juego, cuya traducción aproximada es *juego de rodear*, es controlar una cantidad de territorio mayor a la del oponente. Para dominar un área, debe rodearse con las piedras. Gana quien controle más territorio cuando acabe la partida.
7. En esto seguimos a Latorre (2019). El otro programa que emplea el *machine learning* es Alphazero que, en 2017, venció a Stockfish; Alphazero se entrenó solo contra sí mismo. La imagen, por dura que sea, parece ser la siguiente: ¡hemos creado máquinas que juegan tan bien que ninguno de los mejores jugadores humanos es un rival digno para ellas!
8. La idea de que nuestras criaturas (o nuestros pupilos) lleguen a superarnos es comprensible y razonable. No puedo correr una maratón, pero un atleta entrenado por mí lo hace y logra ganar; no puedo levantar 10 000 kg, pero una máquina diseñada por mí puede hacerlo. En este sentido, los límites de lo que pueden lograr las máquinas diseñadas por nosotros dependen en gran medida de los materiales de los que están hechas, sus diseños, tecnologías, mecanismos y las tareas que deben realizar. Cuando se trata de evaluar las denominadas "máquinas inteligentes", es obligatorio considerar aspectos como la capacidad de cómputo, que ha crecido sin cesar desde la invención de los microprocesadores y que se amplía a medida que se incorporan tecnologías más sofisticadas. Eso explica el éxito de Deep Blue y otros ejemplos similares. No es claro, por ahora, si hay límites empíricos al crecimiento de la capacidad de cómputo o si ya estamos cerca de ellos.
9. Hay dudas razonables sobre la conveniencia de comparar la inteligencia humana con su contrapartida artificial, así como sobre la hipótesis de que las máquinas eventualmente desarrollarán superinteligencia. Dichas dudas se basan en las profundas diferencias entre la forma como los humanos procesamos información, resolvemos problemas y aprendemos, y la manera como se realizan procesos comparables en máquinas dotadas de IA, además de que nuestras capacidades cognitivas sobrepasan de un modo notable la forma algorítmica de aprendizaje de las máquinas (Larson, 2022).

 https://doi.org/10.21500/22563202.7014

> El mito de la inteligencia artificial afirma que su llegada es inevitable y es cuestión de tiempo –pues ya nos hemos embarcado en la ruta que conduce a IA de nivel humano y luego a la súper-inteligencia–. Pero no es así. Dicha ruta solo existe en nuestra imaginación. (p. 1)

Ante estos escenarios posapocalípticos, cabe sospechar que a las generaciones futuras les resultará divertido el temor que expresan sus congéneres actuales acerca de la posibilidad de que las máquinas lleguen a dominarnos, así como a muchos hoy nos resulta gracioso saber que a la mayoría de los británicos, de finales del siglo XIX y comienzos del XX, les causaba pánico usar la luz eléctrica, no solo por posibles electrocuciones –temor mejor fundado–, sino porque consideraban a las lámparas eléctricas como verdaderos enemigos de la belleza: la luz eléctrica era como un horrible detective que revelaba cada arruga y línea de la cara. Naturalmente, estos planteamientos no llevan a concluir que el temor hacia el dominio de las máquinas es una cuestión filosófica baladí. Por el contrario, hay un sentido filosóficamente profundo que está en juego cuando se examina la posibilidad misma de que las máquinas nos subyuguen o puedan llegar a dominarnos, y que tiene que ver con entender que *la dominación de las máquinas* significa nuestra subordinación o sometimiento total a ellas. Ello implicaría la pérdida de nuestra libertad, una consecuencia terrible. La anterior es una suerte de perspectiva apocalíptica y catastrófica: podemos denominar a esta idea *el mito de la rebelión de las máquinas* (Quintanilla, 2017), bien capturado por la imagen de máquinas fuera de nuestro control, matando y sometiendo a los seres humanos de manera horrorosa.

Entre los escenarios distópicos que mejor han representado en el cine el mito de la rebelión y el dominio de las máquinas están las películas *Terminator 2* (1991) de James Cameron y *Matrix* (1999) de las hermanas Wachowski. El argumento de *Terminator 2* es cautivador, no solo por lo fantasioso, sino porque, pese a ello, luce verosímil. Por supuesto, la ficción nos obliga a aceptar que es empíricamente posible que llegue un día en el que las máquinas no estén más bajo nuestro control y que los viajes son posibles en el tiempo.

La guerra entre los seres humanos y las máquinas comienza cuando Skynet, la inteligencia artificial que lidera al ejército de las máquinas y puede controlar el arsenal nuclear de los Estados Unidos por sí sola, envía al pasado a un exterminador para acabar con el líder de la resistencia humana, John Connor, antes de que se convierta en caudillo. El exterminador enviado por Skynet es un T-1000, construido con una "polialeación mimética"; este metal líquido le confiere la propiedad de alterar su figura y adoptar la de cualquier ser u objeto que toque. Aunque ello no aplica para artefactos complejos (armas de fuego y explosivos), el T-1000 puede transformar su cuerpo para que asemeje armas cortopunzantes. En este sentido, toma el aspecto de un policía y persigue al líder de la resistencia.

En el futuro, 2029, John Connor envía al pasado a un T-800 reprogramado, de modo que proteja a la versión joven de sí mismo. Como parte del plan para resistir la rebelión de las máquinas, destruyen el laboratorio de Cyberdyne Systems donde se desarrolla la tecnología de la que surgiría en el futuro Skynet. Como puede advertirse, en *Terminator 2* se plantea un escenario imaginario en el que, como consecuencia del mismo desarrollo tecnológico, emerge una súper inteligencia capaz de salirse de nuestro control: Skynet. Sin embargo, podemos advertir que el control de las máquinas no es total y, gracias a ello, John Connor puede liderar la resistencia contra las máquinas. En esa guerra imaginaria, el ser humano sale victorioso (Falzon, 2015).

El argumento central de *Matrix* es alucinante, guarda muchas similitudes con el de *Terminator* y posee una riqueza notable desde el punto de vista filosófico (Irwin, 2002). La historia tiene lugar, de nuevo, en un futuro distópico. Tras una dura guerra que ha

desembocado en un invierno nuclear, la mayoría de los seres humanos ha sido esclavizada por un poderoso "cerebro" cibernético que ha programado una compleja y flexible simulación computacional interactiva. Los humanos han sido recluidos en cápsulas en las cuales sus cuerpos flotan en una solución que los mantiene vivos y permite su aprovechamiento como fuentes de energía o "baterías humanas". Los cerebros de los humanos esclavizados están conectados a dicha simulación o *matrix*, la cual genera los impulsos eléctricos (estímulos) que producen todas sus sensaciones, experiencias y estados mentales. De este modo, individual y colectivamente, los humanos creen llevar una vida normal que transcurre en la Nueva York de finales del siglo XX, cuando en realidad no tienen ningún contacto real con el mundo exterior ni con otros individuos.

Los pocos seres humanos que no están suspendidos en las cápsulas o que han sido liberados viven en la ciudad de Zion, desde donde planean liberar a otros conectados (Yeffeth, 2003). Morfeo, uno de los habitantes de Zion y personaje central de la historia, basado en una profecía, cree que hay alguien en *Matrix* que puede inclinar la balanza del lado de los humanos en la guerra contra las máquinas: el Elegido. Este personaje es Neo, quien en su vida ilusoria dentro de la Matrix es, de día, un empleado de una inmensa compañía tecnológica, mientras que de noche (o en su "tiempo libre") es un pirata informático. Bajo la suposición de que Neo puede ser el Elegido, Morfeo y Trinity (una miembro destacada del grupo de Morfeo) logran liberar a Neo y le muestran la realidad fuera del sueño universal que genera la *Matrix* para los humanos esclavizados.

La lucha contra las máquinas debe librarse tanto desde afuera como en el ciberespacio, por lo que aquellos liberados deben reconectarse a la simulación interactiva por el lapso que requieren las misiones. Sin embargo, cada vez que se intenta avanzar en la guerra contra las máquinas, hay que enfrentar la persecución de poderosos personajes virtuales creados por ella: los agentes, liderados por Smith, quienes pretenden acceder a los computadores y la infraestructura de Zion. Los agentes capturan a Morfeo y Neo se ve obligado a rescatarlo, arriesgando su vida. Al final de la primera entrega de la historia, Neo se revela como el Elegido y acaba con Smith.

La similitud entre las dos películas es notoria. Tanto en *Terminator* como en *Matrix* se libra una batalla entre los seres humanos y las máquinas, en la que los seres humanos salen vencedores. En ambos casos, hay un derroche asombroso de fantasía y sus respectivos guiones representan como verosímil el desarrollo de la hiperinteligencia, o el *mito de la superinteligencia*, que se asemeja a la controvertida *tesis de la singularidad* (Bostrom, 2014; Kurzweil, 2005).[10] Según esta tesis, en cualquier momento podría ocurrir una explosión súbita e incontrolable de la inteligencia de las máquinas. Lo llamativo es que los escenarios distópicos combinan este mito con la idea correcta –y lógicamente incompatible– del límite al control de las máquinas, que es justo la noción de la que esperamos extraer provecho para el análisis que se propone en este trabajo.

Por lo pronto, basta con advertir que la pregunta acerca de si las máquinas podrían dominarnos en un futuro (cercano o lejano) es lógicamente independiente del asunto de si las máquinas eventualmente podrán pensar de manera autónoma y en el mismo sentido en que lo hacemos los humanos. Esto no quiere decir que los problemas filosóficos relacionados con estas cuestiones no guarden relación alguna. No obstante, es imprescindible reconocer que el examen del primer asunto se puede adelantar, sin tener que dar una respuesta al segundo. De hecho, en los escenarios imaginarios, partiremos de varias suposiciones con base en las cuales se da por sentado que las máquinas adquieren pensamiento autónomo, conciencia y voluntad en un futuro distópico. Pero,

10. Para un análisis crítico de la tesis de la singularidad, ver Zamora Bonilla (2021).

 https://doi.org/10.21500/22563202.7014

como esperamos mostrar, ni siquiera suponiendo que semejante proeza evolutiva de las máquinas sea empíricamente posible, se sigue que la afirmación "las máquinas llegarán a dominarnos" sea lógicamente consistente. En la siguiente sección, ofrecemos una síntesis de las razones empíricas que se pueden esgrimir para mostrar por qué las máquinas no nos dominan ni lo harán en el futuro.

El argumento de la imposibilidad empírica

Como se anunció, es conveniente distinguir entre dos sentidos del término "dominación": uno débil y otro fuerte. De acuerdo con nuestro análisis del sentido débil, la dominación puede referirse a la capacidad de superar en habilidades. Dado que muchas máquinas nos superan en fuerza, velocidad y la capacidad para realizar miles de millones de cálculos con rapidez y precisión, en este sentido, nos dominan. Sin embargo, en el sentido fuerte, el planteamiento de que las máquinas nos dominan debe interpretarse como "las máquinas ejercen un control significativo sobre nuestras vidas" o, de manera más elaborada, las máquinas controlan nuestras acciones determinando por completo el curso de nuestros pensamientos y los estados mentales que conducen a las acciones. Nuestro argumento es que el enunciado "las máquinas llegarán a dominarnos" (en el sentido señalado) parece empíricamente falso y podría ser empíricamente imposible y lógicamente inconsistente. A continuación, exponemos los argumentos que respaldan ambos planteamientos.[11]

El primero de estos argumentos es sencillo y proviene de la imposibilidad misma de producir una hipermáquina todopoderosa con eficiencia absoluta. El argumento se presenta así:

P_1: si las máquinas nos dominan o llegaran a dominarnos (en el sentido fuerte del término), entonces, tienen que ser posibles la eficiencia y el control totales sobre segmentos de la realidad.

P_2: según las leyes de la física, ni la eficiencia ni el control totales son empíricamente posibles.

C: no es empíricamente posible que las máquinas nos dominen.

Las premisas están respaldadas por las leyes de la termodinámica, las cuales establecen que todo trabajo se detiene debido a la disipación de energía y, por lo tanto, ni la eficiencia total ni el control son posibles. Por esta razón, tampoco es factible crear máquinas de movimiento perpetuo,[12] puesto que eliminar los factores que disipan la energía, como la fricción mecánica, es imposible. Instancias promisorias como la máquina de Carnot, cuya eficiencia es del 100 %, son artefactos ideales o imaginarios. Este razonamiento se puede reforzar con otros argumentos perspicaces:

1. El argumento de la voluntad: si las máquinas nos llegaran a dominar, tendrían que desarrollar capacidades volitivas, entre ellas, el *deseo* de autoconservación y reproducción (Diéguez, 2017, p. 61).

11. De acuerdo con el análisis presentado, las máquinas no nos dominan en el sentido fuerte del término, único que es relevante filosóficamente para esta discusión. Dado que las máquinas dotadas de inteligencia artificial requieren un sustrato físico para funcionar (microprocesadores) y una fuente de energía; la demanda de eficiencia total se desprende de las condiciones del argumento. Como lo señaló uno de los revisores anónimos, es importante distinguir entre la implementación efectiva de una máquina y su mera posibilidad. Debido a ello, las máquinas acelerantes de Turing, aunque posibles (en ciertos espacios), por ahora son solo objetos abstractos (Fernández Cuesta, 2023; Nakano, 2018); por lo cual, no cuentan como contraejemplos en esta discusión.
12. Una máquina de movimiento perpetuo tendría, por ejemplo, ruedas que giran eternamente. Serían máquinas que pueden funcionar de manera ininterrumpida y sin la ayuda de energía externa. Como se sabe, la historia está plagada de fraudes cometidos por aquellos que alardearon de poder diseñarlas.

2. El argumento de la autonomía: si las máquinas nos dominaran, tendrían que ser autónomas para satisfacer sus necesidades; ello exigiría, *inter alia*, la capacidad de obtener sus propias fuentes de energía (Diéguez, 2017, p. 62).
3. El argumento de la conciencia: para desarrollar la autonomía y el deseo de dominio que requieren los escenarios de ficción, las máquinas tendrían que ser conscientes.
4. El argumento de las restricciones ecológicas: si las máquinas nos dominaran, tendrían que enfrentarse a una atmósfera rica en oxígeno y en vapor de agua, con el consiguiente efecto degradante y corrosivo sobre sus componentes. La ecología misma no hace favorable las condiciones para una posible progenie rebelde y exitosa (Diéguez, 2017, p. 65).

Es sorprendente que el esfuerzo de imaginación y argumentación dedicado a la provocadora ficción del dominio absoluto de las máquinas sobre los seres humanos ignore estos argumentos y pase por alto la inevitable conclusión científica de que no es empíricamente posible que las máquinas nos dominen sin enfrentar la suposición tácita (y falsa) de que en un escenario de dominio total debe ser realizable la eficiencia total. Los creadores y entusiastas de las distopias, a pesar de su amplio conocimiento de diferentes aspectos de la teoría y la práctica científica que respaldan estas ficciones, han pasado por alto algunas de las condiciones empíricas necesarias para que la historia no solo sea creíble, sino que concuerde con las leyes de la ciencia. Este descuido debilita sus argumentos.

A continuación, mostramos cómo tanto en los escenarios de ficción de *Terminator* y *Matrix* como en otros similares que se podrían proponer es posible demostrar que el planteamiento "las máquinas podrían llegar a dominarnos completa y absolutamente" no solo asume falsamente que la eficiencia total es empíricamente posible, sino que también da lugar a otras inconsistencias.

Escenario imaginario: el control por "grados" de las máquinas

Para ilustrar que el vínculo entre el planteamiento que asume el sentido fuerte del término "dominar" y la suposición de que la eficiencia total es empíricamente posible no es arbitrario, consideremos el siguiente escenario apocalíptico como verdadero, junto con los supuestos necesarios para que funcione y sea plausible.

Supongamos que un día las máquinas logran un desarrollo tan sensacional que ya no nos necesitan para funcionar. En este contexto, adquieren autonomía e independencia para determinar sus propios objetivos y en consecuencia, deciden dominarnos y someternos a su voluntad. Es crucial considerar que, en las nuevas condiciones de este panorama, las máquinas realizan tareas y cumplen objetivos específicos (verbigracia, levantar enormes pesos y realizar miles de millones de cálculos en fracciones de tiempo muy breves). Empero, cuando las máquinas estaban bajo el control de los seres humanos, eran estos quienes determinaban los objetivos de las máquinas.

Ahora, bajo las nuevas condiciones en las que las máquinas se han liberado del control de sus creadores y han establecido sus propios objetivos, asumimos una de sus principales metas: el dominio de los seres humanos. Finalmente, supondremos que las máquinas logran este dominio. Para analizar la conexión entre la tesis de que las máquinas nos dominan y la suposición de la eficiencia total, consideremos que el dominio de las máquinas es de manera gradual y medible en términos porcentuales, de modo que podemos estimar un cumplimiento del 90 % de su objetivo. Esto implicaría que las máquinas nos dominan en un 90 % y nos dominan parcialmente en un 10 %.

Para ilustrar esta idea, supongamos la siguiente condición: X domina a Y si X puede controlar las variables [p, q, r, s], que son esenciales para la autonomía de Y y sobre las

 https://doi.org/10.21500/22563202.7014

cuales Y no tiene control alguno. En tales circunstancias, X controla las variables [p, q, r, s] que corresponden a segmentos de la realidad e involucran funciones de Y. Por ejemplo:

— p: la capacidad de Y para tomar decisiones.

— q: las emociones de Y.

— r: los gustos de Y.

— s: la capacidad de Y para pensar autónomamente.

Como se trata de una dominación gradual, entonces, habría variables que X no controla. Supongamos que se trata de las variables t y w, donde t es respirar y w morir. En *Matrix*, la superinteligencia parece tener control hasta sobre las funciones biológicas básicas de las baterías humanas, pero cuando alguien muere en la simulación interactiva, su cuerpo debe ser descartado, pues ya no puede cumplir su función de generar energía. Hemos elegido al azar el dominio de referencia de las variables [p, q, r, s, …]; sin embargo, para que se pueda mantener la tesis de que las máquinas nos dominan, es claro que deben controlar p. En otras palabras, que X tenga el control de p es una condición necesaria para que Y (los seres humanos) no decidan detener a X. Así las cosas, el control que ejerce X sobre la variable p no puede darse en grados, sino que tiene que ser total, a menos que enriquezcamos la historia con variantes como que X permite que surjan pensamientos de inconformidad en Y, que se formen deseos de liberación, pero que, mediante el control de las demás variables y situaciones fácticas, se mantenga a raya de manera eficaz cualquier intento de rebelión.[13]

Los anteriores planteamientos conducen a la conclusión de que, incluso en el escenario de la dominación por grados, habría que suscribir la premisa de que las máquinas deben poder controlar por completo ciertos segmentos de la realidad y ciertas variables que causan o explican nuestro actuar (*e. g.* la capacidad de toma de decisiones de los seres humanos). Pero el tipo de control que requiere la tesis de la eficiencia total no puede ser parcial y ya mostramos que las leyes de la física son incompatibles con la eficiencia total que acompañaría el supuesto del control total. En consecuencia, que la eficiencia total no sea realizable empíricamente muestra la implausibilidad tanto del control total como del control por grados. No obstante, los partidarios de la tesis de la dominación de las máquinas en sentido fuerte podrían oponer la siguiente objeción: si es posible concebir un escenario de ficción en el que las máquinas desarrollan inteligencia, conciencia y voluntad, ¿por qué no imaginar uno metafísico en el que sea posible la eficiencia total y en el cual tenga lugar la distopia del dominio total de las máquinas? En la siguiente sección analizamos dicho caso y mostramos que conduce a inconsistencias.

Escenario imaginario: el gobierno de las máquinas en el Kepler 452b

Para mostrar que tampoco es lógicamente posible el control total de las máquinas, vamos a imaginar el siguiente escenario apocalíptico. Supongamos que, en un futuro, las máquinas logran –por sí mismas– un desarrollo aún más esplendoroso. Así, no solo se han aventurado hacia un viaje intergaláctico a 1400 años luz, sino que regresan para llevarnos con ellas con el fin de instaurar (en el planeta Kepler 452b) un gobierno de máquinas en el que nos someten a su voluntad. Puesto que en ese planeta no operan

13. Pero esta perspectiva reproduce tan fielmente las emociones y la condición humanas que nos fuerza a adjudicar a las máquinas inteligentes pasiones como maldad, la capacidad de disfrutar del sufrimiento ajeno, entre otras. El problema de asumir que las máquinas pueden desarrollar este tipo de emociones y pasiones no hace más que incrementar la enorme dificultad de aceptar que la inteligencia artificial puede dar el salto a estados que incluyen autoconciencia, voluntad y capacidad de tomar decisiones éticamente cuestionables.

las leyes de la termodinámica, las máquinas encuentran la forma de obtener (o crear) sus propias fuentes de energía. Adicionalmente, puesto que allí la atmósfera no es rica en oxígeno como en la Tierra, las máquinas no enfrentan las restricciones físicas de la corrosión ni los problemas asociados a la fricción. Además, las máquinas evolucionan –bajo las leyes del Kepler 452b– de tal manera que consiguen desarrollar las capacidades volitivas necesarias para *querer* gobernarnos.

Ahora bien, si las máquinas nos dominaran, ello significaría *la pérdida total* de nuestra libertad[14] y autonomía. Con base en ello, podemos suponer que las máquinas tienen el control de las siguientes variables:

— p: la capacidad de tomar decisiones, por lo cual, las máquinas lo hacen por nosotros.

— q: la capacidad de actuar libremente. Aun cuando podemos ejecutar movimientos que son constitutivos de las acciones, sus móviles (resortes o intenciones) están dictados por las máquinas (como en *Matrix*).

— t: la capacidad de pensar por nosotros mismos. Dado que en Kepler 452b las máquinas nos dominan, todos los contenidos de nuestros pensamientos tendrían que ser implantados; es decir, no serían pensamientos libres en el sentido de que no están vinculados natural y causalmente con el mundo de la manera relevante. Todo contenido de pensamiento, en la forma de una proposición, que no guarde un vínculo causal natural y relevante con el mundo externo sería falso o carecería de justificación.

En este sentido, si las máquinas controlan la variable t, ni siquiera sabríamos que somos dominados por ellas, dado que una *conditio sine qua non* para saber (P) (*qua* saber proposicional) es pensar libremente. Si las máquinas nos controlan, no tendríamos contenidos de pensamiento propios, en consecuencia, el pensamiento: "las máquinas nos dominan" tendría que ser un contenido de pensamiento implantado por una máquina. El escenario anterior conduce a la siguiente contradicción.

Supongamos que entre tus contenidos mentales tienes el pensamiento: "las máquinas nos dominan" o (P). Para que (P) sea verdadero, las máquinas tendrían que dominarnos efectivamente. Si es así, (P) parece calificar como pensamiento libre, pues guarda una conexión causal con el mundo, pero para que sea realmente libre no puede ser un pensamiento implantado por una máquina. En ese sentido, si (P) es un pensamiento libre y no ha sido implantado por una máquina, entonces, es falso que las máquinas nos dominan *totalmente*, porque puedes pensar libremente (o puedes pensar lo que quieras).

Sin embargo, todavía es posible imaginar –como en el escenario arriba descrito– una situación en la cual las máquinas, tras dominar por completo todas nuestras funciones cognitivas y volitivas (como en el caso de cerebros en cubetas o de *Matrix*), implantan en nuestras mentes la creencia "las máquinas nos dominan". Desde esta nueva perspectiva, se seguiría necesariamente que la afirmación "las máquinas nos dominan" es verdadera y es falsa. Es verdadera, porque corresponde al estado de cosas (dominación total de las máquinas sobre los humanos) y porque en cuanto contenido de pensamiento implantado por las máquinas, respalda –por sí misma– la afirmación "las máquinas nos dominan". Pero, a la vez, sería falsa (pues podemos formar ese pensamiento por nosotros mismos) o injustificada, porque si aceptamos que todo contenido de pensamiento –en un escenario distópico– es implantado, resulta forzoso concluir que no se da la conexión natural y relevante con el mundo y, en consecuencia, la afirmación "las máquinas nos dominan" no constituye conocimiento en el sentido de la definición tripartita. De este modo, se

14. Entiéndase por "libertad" la capacidad de tomar decisiones y de actuar autónomamente, de acuerdo con la relación tradicional: agencia-conciencia/voluntad-elección. Estas conexiones implican la capacidad de pensar por sí mismos. Por tanto, entiéndase por "pensamiento libre" aquel que guarda una conexión causal natural y relevante con el mundo.

 https://doi.org/10.21500/22563202.7014

configura una inconsistencia epistémica, porque sabemos que (P) y al mismo tiempo no sabemos que (P).

Veámoslo de manera esquemática en la siguiente *reductio*:

1. Supongamos que tienes el pensamiento (P): "las máquinas nos dominan".
2. Para que (P) sea verdadero, las máquinas tendrían que dominarnos. Esto presupone que (P) tiene una conexión causal con el mundo real.
3. Luego, si (P) es un pensamiento libre (que establece la conexión causal relevante con el mundo), no es cierto que las máquinas nos dominen totalmente, pues tienes, por lo menos, un pensamiento libre. En este caso (P) es falso.
4. Pero si las máquinas llegan a dominar incluso tus pensamientos, (P) sería verdadero, *a fortiori*.
5. Luego, (P) es verdadero, porque es un contenido de pensamiento implantado por una máquina, que de paso apoya la afirmación: "las máquinas nos dominan".
6. Adicionalmente, (P) carece de justificación, debido a que en un escenario distópico donde todos los contenidos de pensamiento han sido implantados, no se da la conexión natural y relevante con el mundo real, en consecuencia (P) no constituye conocimiento.

El anterior argumento conduce a una contradicción, específicamente en las premisas tres y cinco dado que si tienes el pensamiento (P) "las máquinas nos dominan", para que (P) sea verdadera, las máquinas tendrían que dominarnos, lo que implica una conexión causal con el mundo. Sin embargo, si (P) guarda una conexión causal con el mundo, entonces, es un pensamiento libre y no es cierto que las máquinas nos dominen totalmente, pues tienes, al menos, un pensamiento libre.

Como se ve, desde una perspectiva parece darse la conexión causal relevante con el mundo mientras que, desde otra, no; y el argumento sugiere que la creencia (pensamiento) es libre (puesto que se da la conexión relevante) y al mismo tiempo no lo es (por haber sido implantado en nuestras mentes). Naturalmente, siempre es posible caracterizar la conexión causal con el mundo de manera tan deflacionaria que el vínculo entre la creencia de que (P) y la agencia de la máquina al implantar dicho pensamiento en nuestras mentes baste para dar por cumplida esta condición. Pero esta movida solo lograría resolver la dificultad para (P), en tanto que la mayoría de nuestras creencias sobre otros asuntos fácticos concernientes al mundo externo, su funcionamiento y nuestras experiencias cotidianas sería falsa, ya que no se da la conexión relevante entre tales creencias y el mundo. En el escenario de *Matrix*, por ejemplo, es claro que ninguna de las creencias de los seres humanos conectados es libre ni verdadera y que el escepticismo sobre el conocimiento es inevitable.

Conclusión

Es fácil advertir que los escenarios distópicos, en los cuales "pensar" equivale a desarrollar nuestras actividades cognitivas de acuerdo con los dictados de una máquina, dan lugar a versiones más sofisticadas y poderosas de los argumentos de Gettier (1963). Como se sabe, estos mantienen las condiciones de creencia y verdad (con algo de manipulación sobre la segunda), pero hacen colapsar la condición de justificación mediante recursos argumentativos que involucran coincidencias bizarras o plantean que se satisfacen conjuntamente las condiciones de la definición tripartita de "conocimiento" aunque, pese a ello, no hay conocimiento, lo que mostraría la insuficiencia de dicha definición. Por el contrario, los argumentos que dependen de los escenarios distópicos cuestionan directa y fatalmente la condición de verdad, puesto que, si las máquinas implantan creencias

en nuestra mente, el análisis muestra que algunas de esas creencias (como la de que las máquinas nos dominan), siendo verdaderas, *prima facie*, no se deberían juzgar como verdaderas, aun cuando guarden la relación apropiada con el estado de cosas. El fallo radica en que la creencia es verdadera, pero no se da la conexión causal natural y relevante entre la creencia y el mundo. Para apreciar mejor esta nueva clase de argumentos de Gettier,[15] examinemos el siguiente esquema:

S sabe que P ("las máquinas ejercen un dominio absoluto sobre los seres humanos") si:

(i) S tiene la creencia de que P.

(ii) La creencia de que P ha sido formada mediante una conexión causal relevante y natural con el mundo, y dicha conexión es esencial para determinar correctamente la verdad o falsedad de la creencia (es decir, la interacción ordinaria que tenemos con el mundo provoca la creencia en cuestión, y en condiciones normales formamos las creencias correctas).

(iii) P es verdadera.

(iv) La creencia de que P está satisfactoriamente justificada.

A la luz de este análisis, resulta obligatorio reconocer que los fallos de (ii) afectan fatalmente a los de (iii) y debemos concluir que, en aquellas circunstancias en las cuales la creencia de que P ha sido implantada por una máquina, S no sabe que P, porque P parece ser simultáneamente verdadera y falsa, además de no tener la relación causal y natural relevante con el estado de cosas (indispensable para predicar la verdad).

En lenguaje cotidiano, resulta forzoso concluir que la posibilidad empírica de que las máquinas logren un control absoluto sobre los seres humanos (incluyendo el de sus mentes) parece irrealizable, por las limitaciones que imponen las leyes físicas. Adicionalmente, la posibilidad lógica de dicho dominio resulta amenazada por las inconsistencias (lógicas y epistémicas) derivadas de las premisas en la caracterización de un evento tal, bajo el análisis presentado. A pesar de que estamos frente a un escenario concebible, si se diera el hecho hipotético de un domino absoluto de las máquinas sobre los seres humanos (como ocurre en *Matrix*), no podríamos saberlo (ni en el sentido clásico de la definición tripartita ni en el de la definición alternativa propuesto) y, finalmente, queda claro que el temor (omnipresente en la literatura y el cine de ciencia ficción) a que las máquinas lleguen a ejercer ese tipo de dominio no está bien fundamentado filosóficamente.

Referencias

Asimov, I. (1950). *I, Robot.* Doubleday.

Boden, M. (2016). *AI: Its nature and future.* Oxford University Press.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford University Press.

Cameron, J. (Dir.). (1991). *Terminator 2* [Película]. Carolco Pictures; Pacific Western Productions; Lightstorm Entertainment; StudioCanal.

Cassini, A. (1997). Equivalencia empírica y subdeterminación en las teorías físicas. *Crítica, Revista Hispanoamericana de Filosofía*, *24*(86), 3-51. https://doi.org/10.22201/iifs.18704905e.1997.1062

15. A diferencia de los argumentos de Gettier clásicos, en los cuales el planteo de conocimiento colapsa debido a ambigüedades en las descripciones definidas, coincidencias extrañas u operaciones lógicas aparentemente lícitas (que en realidad no son correctas), este tipo de argumento depende de la caracterización del rol de la realidad (mundo) en el proceso de formación de creencias; es decir, de la conexión causal relevante entre el mundo y el proceso de formación de nuestras creencias. Para más detalles sobre el análisis de los contraejemplos de Gettier, se sugiere el trabajo de García Duque (2007).

Chalmers, D. (2002). Does conceivability entail possibility? En T. S. Gendler y J. Hawthorne (Eds.), *Conceivability and possibility* (pp. 145-300). Oxford University Press.

Descartes, R. (2009). *Meditaciones metafísicas* (P. Pavesi, ed.). Prometeo.

Diéguez, A. (2017). *Transhumanismo: la búsqueda tecnológica del mejoramiento humano*. Herder.

Falzon, C. (2015). *Philosophy goes to the movies: An introduction to philosophy.* Routledge.

Fernández Cuesta, J. A. (2023). ¿Existen las máquinas aceleradas de Turing? Paradojas y posibilidades lógicas. *Techno Review: International Technology Science and Society Review*, *13*(1), 49-74.

Freeland, C., y Wartenberg, T. (1995). *Philosophy and film.* Routledge.

Galileo. (2004). *Diálogos acerca de dos nuevas ciencias*. Losada.

García Duque, C. E. (2007). Casos Gettier y razonadores normales. *Ideas y Valores*, *56*(135), 77-88. https://revistas.unal.edu.co/index.php/idval/article/view/1140

Gendler, T. S., y Hawthorne, J. (2002). Introduction: Conceivability and possibility. En T. S. Gendler y J. Hawthorne (Eds.), *Conceivability and possibility* (pp. 1-10). Oxford University Press.

Gettier, E. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121-123. https://doi.org/10.2307/3326922

Irwin, W. (Ed). (2002). *The Matrix and philosophy: Welcome to the dessert of the real.* Open Court.

Kirk, R. (2023). Zombies. En E. N. Zalta y U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy.* https://plato.stanford.edu/entries/zombies/

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology.* Penguin.

Larson, E. J. (2022). *The myth of artificial intelligence: Why computers can't think the way we do.* Belnak Press of Harvard University Press.

Latorre, J. I. (2019). *Ética para máquinas.* Ariel.

Nakano, A. L. (2018). Máquinas de Zenão e a distinção entre cálculo e experimento. En *Filosofía e historia de la ciencia en el Cono Sur* (pp. 178-186). AFHIC.

Neri, H., y Pessoa Jr., O. (2018). Science without consciousness. En *Filosofía e historia de la ciencia en el Cono Sur* (pp. 149-155). AFHIC.

Ortega y Gasset, J. (1947). Meditación de la técnica. En *Obras completas* (pp. 317-275). Revista de Occidente.

Oyarzún, P. (1999). Epicuro: Carta a Meneceo. *Onomázein*, *4*, 403-425. https://doi.org/10.7764/onomazein.4.22

Quintanilla, M. Á. (2017). *Tecnología: un enfoque filosófico y otros ensayos de filosofía de la tecnología.* FCE.

Reuleaux, F. (1875). *Theoretische Kinematik. Grundzüge einer Theorie des Maschinenwesens.* Friedrich Vieweg und Sohn.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*, 433-460. https://doi.org/10.1093/mind/LIX.236.433

Wachowski, L., y Wachowski, L. (Dirs.). (1999). *Matrix* [Película]. Village Roadshow Pictures; Silver Pictures.

Yeffeth, G. (Ed.). (2003). *Taking the red pill: Science, philosophy and the religion in the Matrix.* Benbella Books.

Zamora Bonilla, J. (2021). *Contra apocalípticos: ecologismo, animalismo, posthumanismo*. Schackleton Books.