ELISA SIMÓ SOLER
ELOY PEÑA ASENSIO
(Coordinación)

DEFENSA PLANETARIA

Autoría:

Alba Soriano Arnanz
Albert Rimola
Alberto Coronel Tarancón
Anna Garcia Hom
Catia Faria
Elisa Simó Soler
Elisa Celia González Ferreiro
Eloy Peña Asensio
Jordi Solé i Ollé
José Ignacio Robles Sánchez
Josep Maria Trigo-Rodríguez
Juan Manuel de Faramiñán Gilbert
Juan Miguel Sánchez Lozano
Julia de León
Nadjejda Vicente Cabañas
Ramon J. Moles Plaza

Dykinson, S. L.

Este libro ha sido sometido a evaluación por parte de nuestro Consejo Editorial
Para mayor información, véase www.dykinson.com/quienes_somos

Editorial DYKINSON, S.L. Meléndez Valdés, 61 - 28015 Madrid
Teléfono (+34) 91 544 28 46 - (+34) 91 544 28 69
e-mail: info@dykinson.com
http://www.dykinson.es
http://www.dykinson.com

ISBN: 978-84-1122-441-3
Depósito Legal: M-31318-2023
DOI: 10.14679/2271

ISBN electrónico: 978-84-1170-831-9

Maquetación:
german.balaguer@gmail.com

CAPÍTULO 12. LA INTELIGENCIA ARTIFICIAL Y OTROS RIESGOS CATASTRÓFICOS O EXISTENCIALES

Alba Soriano Arnanz[1]
Profesora Ayudante Doctora de Derecho Administrativo, Universitat de València

DOI: 10.14679/2284

Sumario: 1. INTRODUCCIÓN. 2. ESTADO DE LA CUESTIÓN. 2.1. La creciente penetración de la IA. 2.2. Algunos riesgos asociados a los usos actuales de la IA. 3. LA IA COMO NUEVO RIESGO CATASTRÓFICO: PARALELISMOS CON OTRAS GRANDES CRISIS. 3.1. Escenarios catastróficos. 3.2. Lecciones derivadas del estudio de otras grandes catástrofes. 4. LA IA COMO HERRAMIENTA DE MITIGACIÓN DE OTROS RIESGOS EXISTENCIALES. 4.1. Prevención y gestión del cambio climático. 4.2. Prevención y gestión del impacto de un asteroide contra la Tierra. 5. CONCLUSIONES.

1. INTRODUCCIÓN

La Inteligencia Artificial (IA) ha revolucionado la forma en que las empresas y los gobiernos toman decisiones. La capacidad de procesar grandes cantidades de datos y de aprender de ellos permite tomar decisiones más informadas y rápidas, convirtiéndose en una herramienta cada vez más valiosa para la toma de decisiones en una amplia gama de sectores. Sin embargo, a medida que los sistemas basados en IA proliferan, surgen nuevos paradigmas que conllevan retos regulatorios.

Uno de los mayores desafíos es la evaluación de riesgos directos de la IA y riesgos asociados a su delegación. La IA es un sistema que está diseñado para aprender por sí mismo de los datos que se le proporcionan, lo que significa que si los datos que se utilizan para entrenar a la IA están sesgados o son incompletos, la IA puede tomar decisiones erróneas o incluso discriminatorias. Esto plantea preguntas sobre la representatividad y la soberanía: ¿quién es responsable de las decisiones que toma la IA? y ¿cómo podemos garantizar que estas decisiones sean justas y equitativas?

[1] Doctora en Derecho por la Universidad de Valencia y Profesora Ayudante Doctora en el Departamento de Derecho Administrativo de la misma universidad. Su línea principal de investigación se centra en la regulación de las nuevas tecnologías desde la perspectiva de la protección de los derechos fundamentales.

Este dilema es particularmente relevante debido a la cada vez más clara tendencia en expansión de los sistemas de IA. La capacidad de la IA supera con creces la capacidad humana, reportando innegables beneficios, pero cada vez que desarrollamos nuevos sistemas, surgen nuevos riesgos. La IA es, por un lado, una herramienta salvadora que puede ayudarnos a resolver problemas difíciles. Sin embargo, también es nuestra potencial perdición, ya que puede conducir a resultados no deseados y desafíos imprevistos que podrían llegar a adquirir una escala global.

Las aplicaciones prácticas de la IA son numerosas y variadas. Se utiliza en la atención médica para asistir en la toma de decisiones sobre el diagnóstico y el tratamiento, en la industria para mejorar la eficiencia de los procesos y en la seguridad para detectar y prevenir daños. Pero estas aplicaciones prácticas también plantean preguntas importantes sobre las políticas públicas en relación a la IA. ¿Cómo podemos garantizar que la IA se utilice de manera ética y responsable, y cómo podemos proteger a la ciudadanía de las consecuencias negativas de la IA? ¿Podría descontrolarse una IA muy avanzada? Y si es así, ¿qué medidas preventivas deberíamos adoptar?

A grandes rasgos, los riesgos existenciales son aquellos que pueden terminar con la humanidad tal y como la conocemos, bien por producir su aniquilación absoluta o bien por provocar tal destrucción y cambio que la civilización deje de existir en su forma actual (**Bostrom 2002**). En este sentido, la Ley de Gestión de Riesgos Catastróficos Globales, recientemente aprobada por el Congreso de los EE. UU.[2], define los riesgos existenciales como aquellos que podrían provocar la extinción humana, al tiempo que los riesgos catastróficos serían los que podrían generar acontecimientos que dañasen o hiciesen «retroceder de forma significativa a la civilización humana a escala mundial».

Con el desarrollo acelerado de la IA en las últimas décadas, se ha comenzado a explorar su impacto en la seguridad global y su potencial para crear riesgos catastróficos o existenciales. Uno de los mayores temores es la posibilidad de que la IA llegue a desarrollar una «superinteligencia» que supere ampliamente la capacidad humana de comprensión y control, lo que podría llevar a la creación de un agente artificial que no esté alineado con los valores humanos y que pueda causar daño a gran escala (**Bostrom 2014**).

Este escenario se conoce como «singularidad tecnológica» y plantea la posibilidad de que la IA pueda evolucionar fuera de nuestro control mejorándose a sí misma, llegando a un punto en el que la humanidad ya no tenga la capacidad de entender ni predecir ni bloquear sus acciones. Una «superinteligencia» de este tipo podría tomar decisiones que pongan en riesgo la existencia de la humanidad, ya sea intencionalmente o por error.

Si bien es cierto que todavía parece que nos encontramos lejos de crear esa «superinteligencia», también lo es que los sistemas ya existentes de IA han demostrado generar riesgos significativos para los derechos de las personas. Por esta razón, estimamos conveniente partir de la situación actual, en la que, de manera cada vez más

[2] Global Catastrophic Risk Management Act (2022): https://www.congress.gov/117/bills/hr7776/BILLS-117hr7776enr.pdf#page=1290.

habitual, se automatizan diferentes tipos de procesos de gestión y toma de decisiones con el objetivo de incrementar su eficiencia. Así, la segunda sección de este trabajo establece la forma en la que la IA ha ido penetrando en todos los ámbitos y los peligros que ya entraña, para luego pasar en la tercera sección a abordar los posibles riesgos catastróficos o existenciales que la potencial aparición de una IA muy avanzada podría tener para la humanidad.

Finalmente, la cuarta sección del trabajo se centra en analizar los paralelismos entre la IA y otros riesgos que podrían llegar a provocar daños extremos o la aniquilación de la humanidad. Asimismo, consideramos que el estudio anticipatorio de riesgos de la IA es particularmente relevante por cuanto que, a pesar de presentar riesgos muy significativos derivados fundamentalmente de la forma en la que los seres humanos decidimos utilizarla, si se emplea de manera correcta, puede contribuir a prevenir y gestionar otras situaciones de catástrofe como el cambio climático antropogénico o el impacto de un asteroide contra la Tierra. Estas cuestiones se abordan también en la sección cuarta del presente capítulo.

2. ESTADO DE LA CUESTIÓN

2.1. La creciente penetración de la IA

La IA es un término que hace referencia a la capacidad de las máquinas de realizar procesos típicos de la cognición humana. Es un sistema conformado por algoritmos, es decir, por un conjunto de instrucciones y operaciones matemáticas de índole mayoritariamente estadística, que pretenden minimizar el coste de la tarea que se le asigne.

La versión más reciente del artículo 3.1 de la Propuesta de Reglamento de Inteligencia Artificial de la Unión Europea, define los sistemas de IA en los siguientes términos:

> «Un sistema diseñado para operar con un cierto nivel de autonomía y que, basándose en datos e insumos proporcionados por máquinas y/o personas, infiere cómo lograr un conjunto determinado de objetivos definidos por el ser humano utilizando enfoques basados en el aprendizaje de máquinas y/o en la lógica y el conocimiento, y produce resultados generados por el sistema, como contenidos (sistemas de IA generativa), predicciones, recomendaciones o decisiones, que influyen en los entornos con los que interactúa el sistema de IA»[3].

[3] Traducción propia realizada a partir de la versión en inglés, que es la única publicada hasta la fecha: «'artificial intelligence system' (AI system) means a system that is designed to operate with a certain level of autonomy and that, based on machine and/or human-provided data and inputs, infers how to achieve a given set of human-defined objectives using machine learning and/or logic- and knowledge based approaches, and produces system-generated outputs such as content (generative AI systems), predictions, recommendations or decisions, influencing the environments with which the AI system interacts».

A menudo se hace referencia a dos tipos de IA: la débil o limitada y la fuerte o general. En este apartado nos centraremos en la primera, aquella que considera que el comportamiento humano puede ser utilizado como modelo para entrenar algoritmos capaces de resolver problemas complejos. Sin embargo, es importante tener en cuenta que los algoritmos únicamente realizan tareas específicas en función de su programación, sin llegar a comprender o deducir el significado detrás de la orden que se les ha proporcionado (**Simó Soler y Rosso 2022**).

Una técnica muy común utilizada en la elaboración de modelos de IA es el aprendizaje automático (*machine learning* en inglés), donde una serie de algoritmos utilizan datos para aprender y mejorar analizando la información y actualizándose a medida que se van retroalimentando de los efectos (aciertos y errores) de las decisiones previamente tomadas, un proceso conocido como «entrenamiento». De esta manera, la IA puede adaptarse a nuevas situaciones y aumentar su capacidad de resolver problemas con el tiempo.

El desarrollo de la IA se viene dando en los últimos años de una forma muy diferente a la que se preveía por el imaginario colectivo. La implementación de nuevas herramientas de IA va teniendo lugar de manera bastante progresiva, incorporándose de forma orgánica a diferentes procesos productivos o decisorios. La IA no tiene generalmente la apariencia de robots humanoides sino la forma de funcionalidades integradas en dispositivos, como un altavoz (asistentes de voz) o un *software* que da órdenes a la maquinaria en procesos productivos. Los seres humanos generalmente aceptamos acríticamente su utilización al considerar estos elementos como meros instrumentos que facilitan el trabajo y aumentan la eficiencia de los procesos (pensemos en Google Maps o Siri). Esta permeación de la IA a nuestras vidas coincide además con un momento en el que la ideología dominante tiende a la priorización de la eficiencia definida desde la perspectiva y con el objetivo principal de lograr la mayor ganancia o menor pérdida económica directamente derivada de la actividad realizada. La mayoría de estos sistemas de IA actuales son de tipo débil o limitado, ya que están diseñados para realizar tareas concretas como traducción de idiomas, identificación de objetos en imágenes, recomendación de contenidos, búsqueda de soluciones óptimas o predicción de resultados.

La IA se emplea, cada vez más, en actividades presentes en los departamentos de recursos humanos, en la concesión de préstamos bancarios o la gestión de servicios públicos. Estos sistemas ofrecen la posibilidad de analizar grandes cantidades de información relativa a las personas que están evaluando y formular predicciones mucho más precisas que las que podría realizar cualquier ser humano referidas, por ejemplo, a la adecuación de una persona para un puesto de trabajo, la probabilidad de que incurra en un impago de un crédito o a la asignación más eficiente de recursos públicos.

2.2. Algunos riesgos asociados a los usos actuales de la IA

Si bien es cierto que el uso de la IA en todos los ámbitos produce importantes ventajas y mejora enormemente la eficiencia de los procesos decisorios, no podemos

negar que su aplicación no se encuentra ausente de riesgos que ya han sido señalados en numerosas ocasiones por la doctrina (**Barona Vilar 2021**; **Mittelstadt et al. 2016**).

Los daños y riesgos generados por el empleo de la IA vienen dados, en no pocos casos, por las decisiones humanas que se toman en relación con la forma y objetivos para los que se utilizan estos sistemas o por una excesiva confianza en su buen funcionamiento sin establecer los controles y salvaguardas necesarios para asegurar el respeto al ordenamiento jurídico, como veremos a lo largo de los siguientes párrafos. Esta puntualización es relevante dado que no conviene caer en discursos que criminalicen la propia existencia o el uso de la IA. Los sistemas de IA son una herramienta muy útil en la mayor parte de contextos en los que se emplean. Sin embargo, su capacidad para analizar grandes cantidades de información personal hace que debamos tener un especial cuidado en cuanto a la privacidad de los datos y las implicaciones de su utilización.

Por ejemplo, si se opta por utilizar un sistema de IA para detectar fraudes en programas de ayuda social, es posible que se contribuya a perpetuar estereotipos negativos sobre personas en situaciones socioeconómicas particularmente vulnerables, como la creencia de que son responsables de su propia situación o que buscan aprovecharse del Estado y sus subvenciones (**Alston 2019**; **Ranchordás y Schuurmans 2020**). En este caso, al emplear sistemas de IA se reconoce la existencia de un problema y se legitima la criminalización y señalamiento de las personas vulnerables.

Asimismo, si se confía ciegamente en los resultados de los sistemas de IA y no se establecen los controles oportunos, pueden llegar a multiplicarse las situaciones de desventaja o vulnerabilidad de determinados grupos (**Akselrod 2021**; **Gerards y Xenidis 2021**). Por ejemplo, imaginemos que la actuación de todo un departamento de recursos humanos de una gran empresa se sustituye por una IA sesgada que ha aprendido que deben rechazarse las candidaturas presentadas por mujeres tanto para trabajar en la empresa como para ascender. Evidentemente, aunque algunas de las personas del departamento de recursos humanos tengan prejuicios y hayan tomado en el pasado decisiones discriminatorias contra las mujeres, incluso sin ser conscientes de ello, también habrá casos en los que se haya contratado o concedido ascensos a mujeres. Sin embargo, si el sistema incorpora un sesgo contra las mujeres y, por tanto, entiende que su función es tratar de maximizar este objetivo, esto es, contratar al menor número de mujeres posible, la sustitución de muchos seres humanos por un único sistema sesgado capaz de procesar por sí solo todas las solicitudes de empleo y promoción puede derivar en que se magnifique la situación de desventaja sufrida por las mujeres.

Esto no supone una enmienda a estos sistemas, ya que es posible utilizarlos si se establecen controles adecuados para detectar y corregir sesgos, siendo crucial tener en cuenta el riesgo que implica no implementar dichos mecanismos de supervisión.

Cabe también destacar el riesgo de que se empleen sistemas de IA con el objetivo de influir en procesos democráticos. Así ocurrió con la creación de perfiles precisos y simplificados con características bien definidas de grupos objetivos con la intención de manipular a votantes en las elecciones de los EE. UU. del año 2016 (**Weller 2019**). Asimismo, herramientas potentes de generación de perfiles pueden utilizarse también

en el contexto de la publicidad dirigida y depredadora en la que se lanzan mensajes específicamente diseñados a personas en situaciones de particular vulnerabilidad con el objetivo de que adquieran productos financieros tóxicos (**O'Neil 2016**; **Yeung 2017**).

Por otra parte, teniendo en cuenta que la IA se está desarrollando y empleando fundamentalmente por parte del sector privado, estamos asistiendo, en la actualidad, a un desplazamiento creciente del poder desde instituciones públicas democráticas a empresas privadas que actúan, *de facto*, como reguladores monopolísticos u oligopolísticos en el mercado tecnológico (**Nadler y Cicilline 2020**).

Debemos destacar también los riesgos para el derecho fundamental a la protección de datos personales que genera el uso de estas herramientas, entre otras cuestiones, porque son capaces de inferir de manera precisa información personal incluso cuando esta no quiere ser compartida. Esta es una cuestión particularmente relevante por cuanto, hasta la fecha, no hay un pronunciamiento claro sobre si los datos personales inferidos se encuentran amparados por la normativa en materia de protección de datos y todo apunta a que, en realidad, no lo están. En relación con esta cuestión, el uso de la IA puede generar daños para la autonomía y libertad de las personas que pierden la capacidad de decidir cómo se presentan al mundo, pues la información que de ellas se dispone depende, cada vez más, de los perfiles elaborados por programas informáticos (**Wachter y Mittelstadt 2019**).

La presencia creciente de las aplicaciones de IA se encuentra, sin duda, entre los grandes retos de nuestra sociedad actual. Sin haber llegado a cumplirse aquellas predicciones futuristas en las que se planteaba la sustitución plena de seres humanos por robots humanoides, en la última década hemos podido ver cómo se han creado programas informáticos capaces de asimilar y procesar información a una velocidad inalcanzable para cualquier cerebro humano. La penetración de sistemas automatizados en toda clase de contextos y, en particular, en aquellos en los que se toman decisiones que afectan de manera directa a la vida de las personas, ha generado la necesidad de regular su uso, sobre todo, al detectarse los significativos riesgos que la IA puede llegar a generar para los derechos de la ciudadanía, así como para otros valores e intereses que resulta esencial proteger en el marco de cualquier Estado democrático.

3. LA IA COMO NUEVO RIESGO CATASTRÓFICO: PARALELISMOS CON OTRAS GRANDES CRISIS

El desarrollo cada más avanzado de los sistemas de IA, con sus riesgos y beneficios asociados, puede conducir a un descontrol de estas tecnologías hasta llegar, incluso sin ser conscientes de ello, al punto de inflexión en el que la IA no solamente amenace algunos de los principios básicos de nuestras sociedades, sino que ponga en peligro la propia existencia de la humanidad. Esto es así, sobre todo si tenemos en cuenta que, hasta la fecha, las estrategias de regulación han trasladado la carga de protección de

sus propios derechos a los individuos (este es el caso de la regulación en materia de protección de datos que se basa fundamentalmente en el consentimiento)[4], o directamente han confiado el control del cumplimiento normativo de los sistemas de IA a las propias entidades que los desarrollan, siendo esta la línea que ha adoptado la propuesta de reglamento de IA de la Unión Europea (**Soriano Arnanz 2021c**).

Por ello, más allá de los problemas que ya genera el uso de la IA y frente a los que se pueden y deben ir adoptando soluciones regulatorias, no podemos dejar de lado, en particular en un trabajo en el que se vincula la IA con la defensa planetaria, el papel que la IA puede tener como origen de posibles catástrofes. En esta sección establecemos, por una parte, algunos escenarios de desarrollo de la IA hasta el punto de producir un resultado catastrófico para la humanidad y, por otra, los paralelismos y la vinculación entre la IA y otras fuentes de catástrofes como el cambio climático o el futurible impacto de un asteroide contra la Tierra.

3.1. Escenarios catastróficos

Son múltiples los escenarios que se barajan cuando hablamos de los potenciales efectos catastróficos de la IA, entre los que podemos referirnos, por ejemplo, a la destrucción total de la humanidad o al sometimiento de los seres humanos a una «superinteligencia» en un régimen de esclavitud (**Turchin y Denkenberger 2020**).

3.1.1. Riesgos de una inteligencia artificial limitada

La mayoría de discusiones relativas a los posibles riesgos catastróficos derivados del uso de la IA se centran en los efectos del desarrollo de una inteligencia artificial fuerte, también conocida como inteligencia artificial general. El concepto de inteligencia artificial fuerte se refiere a «sistemas de IA que posean un grado razonable de autocomprensión y autocontrol autónomo, y que tengan la capacidad de resolver una variedad de problemas complejos en una variedad de contextos, y de aprender a resolver nuevos problemas que desconocían en el momento de su creación» (**Goertzel y Pennachin 2007**).

La IA fuerte o general se distingue así de los sistemas de IA limitada en que estos últimos se crean con el objetivo de dar solución a problemas muy concretos como, por ejemplo, jugar una partida de ajedrez o detectar el riesgo de que una persona convaleciente en postoperatorio desarrolle un caso de sepsis, pero no son capaces de realizar funciones diferentes de aquellas para las que se han desarrollado. Ahora bien, el potencial limitado de estos sistemas no implica que debamos despreciar los daños que pueden producir.

[4] Es abundante la doctrina que profundiza en las razones por las que el consentimiento ha fracasado como herramienta de protección del derecho fundamental a la protección de datos personales. Ver, por ejemplo, **Huergo Lora** (**2020**).

Al abordar los riesgos de los programas de IA con funciones limitadas debemos referirnos, necesariamente, a la posibilidad de que, de manera voluntaria o involuntaria, se creen virus informáticos que dañen infraestructuras clave o incluso de manera directa a seres humanos. Estos riesgos se encuentran íntimamente relacionados con la creciente automatización de muchas acciones y procesos y la expansión de lo que se ha denominado como «internet de las cosas» (**Nord, Koohang y Paliszkiewicz 2019**; **Turchin y Denkenberger 2020**).

La cada vez mayor dependencia que nuestras sociedades tienen de elementos que pueden ser manipulados por virus informáticos nos expone a amenazas de diverso tipo. Por ejemplo, un ataque informático a plantas químicas podría derivar en que se liberasen sustancias con muy elevados niveles de toxicidad de forma descontrolada. Asimismo, la creciente penetración de la domótica en los hogares de las personas implica que existe una vulnerabilidad significativa frente a ataques a los sistemas electrónicos, pudiendo producirse incendios u otros daños (**Abomhara y Køien 2015**). Un ataque masivo a las redes eléctricas podría frenar, entre otras cuestiones, la producción de alimentos y la provisión de agua (**Cole et al. 2016**). Los coches autónomos representan también un potencial riesgo evidente, ya que su programación podría ser manipulada con el objetivo de modificar sus comandos y provocar acciones mortales.

Otros de los riesgos sobre los que se ha llamado la atención son aquellos que se refieren al uso de la IA por agentes particularmente poderosos con el objetivo de vigilar y controlar a toda la humanidad. En este contexto, la IA serviría como herramienta empleada por seres humanos en un régimen de dictadura global y nos hallaríamos en una sociedad similar a la descrita por George Orwell en «1984» (**Turchin y Denkenberger 2020**). Sin embargo, lo cierto es que este escenario no parece del todo probable si tenemos en cuenta que, en realidad, el desarrollo de la IA permite ejercer el control sobre las personas de una manera mucho más sutil y pacífica. Así, a través de las redes sociales y plataformas de servicios en línea se puede influir de manera significativa en el comportamiento humano, teniendo estas un potencial de manipulación que previsiblemente irá creciendo de manera exponencial durante los próximos años.

Una preocupación adicional en relación con los efectos negativos de una IA limitada es el desarrollo potencial de armas que tengan un gran poder destructivo. Si no se establecen medidas adecuadas, estas armas podrían ser utilizadas por individuos o grupos para llevar a cabo la eliminación selectiva de una parte de la población. Por ejemplo, enjambres de drones con capacidad para matar a seres humanos, equipados con tecnología de reconocimiento facial para identificar a miembros específicos de la población, podrían ser utilizados para llevar a cabo genocidios. Es crucial considerar estas posibilidades y establecer regulaciones para prevenir estos escenarios catastróficos (**Maas et al. 2023**).

Los riesgos aquí descritos son solo unos pocos de todos los escenarios catastróficos que, con mayor o menor probabilidad, podrían llegar a originarse a partir del uso de sistemas de IA limitados. En todas estas situaciones no estamos hablando de la narrativa comúnmente instalada en el imaginario social en la cual se produce un desarrollo tal de la IA que esta adquiere la capacidad de controlar a los seres humanos, sino del resultado

de decisiones o de errores humanos. Asimismo, podríamos encontrarnos con eventos catastróficos que sean una consecuencia combinada de una decisión humana que se toma con el objetivo de causar una destrucción limitada pero que, como resultado de un error en la programación o en el desarrollo de la máquina, terminen por provocar una catástrofe a gran escala. Por ejemplo, si en el caso de los drones de guerra que hubiesen recibido la orden de matar a grupos concretos de personas, el sistema, fruto de un error o casualidad, acabase reprogramando como objetivo la exterminación de toda la humanidad.

3.1.2. El desarrollo de una «superinteligencia»

El punto de inflexión en la evolución de la IA tendría lugar en el momento en el que existiese un sistema capaz de mejorarse de manera autónoma y suficientemente significativa como para llegar a autotransformarse resultando en una expansión de sus capacidades. En este sentido, resulta necesario diferenciar entre la capacidad de autoaprendizaje y la posibilidad de mejorarse de manera autónoma de los sistemas de IA. En términos muy generales podríamos hablar, ya hoy en día, de la capacidad de los sistemas de IA para «automejorarse» en la medida en la que los sistemas de *machine* y *deep learning* son capaces de autoevaluar los resultados que han ofrecido y adaptar sus parámetros de análisis para perfeccionar su nivel de precisión. Ahora bien, para hablar de verdadera «automejora», aquella que sería el primer paso en el desarrollo de una IA general y que podría derivar en una pérdida de control por parte de los seres humanos, debemos referirnos a sistemas que tengan la capacidad de modificar su propio diseño para convertirse en algo distinto (**Kumar 2019**).

Es importante señalar que este trabajo se enfoca únicamente en los posibles riesgos asociados a la «superinteligencia» desarrollada por una IA que no imita completamente el cerebro humano. Esto se debe a que existen mayores peligros cuando una IA sigue procesos de razonamiento diferentes a los de los seres humanos. La dificultad para comprender su funcionamiento interno y el hecho de que no se rija por los mismos principios que el cerebro humano aumenta la probabilidad de que estos sistemas se descontrolen y produzcan efectos catastróficos para la humanidad (**Hilton 2022**).

3.1.3. Velocidad del despegue de la «superinteligencia»

Uno de los elementos más relevantes relacionados con la generación de una IA general es si el despegue será lento o rápido. Esto es, la velocidad a la que se irán produciendo las nuevas y mejoradas versiones del sistema hasta que los seres humanos no puedan controlarla y, en el peor de los casos, que el sistema incluso pueda terminar con la humanidad. Sin llegar a adentrarnos en estas discusiones, ya que no es realmente posible predecir qué tipo de evolución de la IA general tendría lugar, estimamos conveniente referirnos a los especiales riesgos de un despegue rápido, puesto que dificultará

que puedan adoptarse mecanismos para hacer frente a los daños ocasionados (**Bostrom 2014**).

Otra de las razones por las que la velocidad del despegue es también relevante es que, si una sola IA consigue lo que se conoce como «ventaja estratégica decisiva», es decir, adquiere un desarrollo suficiente como para situarse muy por delante de cualquier otra IA que esté en proceso de convertirse en general o fuerte, esto podría derivar en un escenario en el que aparezca lo que en inglés se ha denominado como «AI singleton» y que podríamos traducir al castellano como «IA singular». Este concepto se refiere a un escenario hipotético en el que una IA superinteligente se vuelve tan poderosa e influyente que deviene en la única entidad dominante en el planeta, efectivamente transformándose en un «singleton» o IA singular que podría incluso llegar a desactivar o ejercer control sobre otras IA que pudiesen situarse como posibles competidoras. Cabe destacar que este escenario no necesariamente tendría que ser catastrófico, ya que existe la posibilidad de que esta IA singular gobernase el mundo de manera benévola (**Bostrom 2014**).

Sin embargo, no es en absoluto descartable que se puedan llegar a dar escenarios catastróficos como consecuencia directa del desarrollo de dos o más IA de potencia igual o similar en la medida en la que la lucha entre estas por el poder podría derivar en una aniquilación absoluta de la humanidad. En una hipotética situación en la que se creen dos o más IA y alcancen un nivel de inteligencia y capacidad de aprendizaje suficiente, podrían comenzar a competir por recursos y poder en su afán por lograr sus objetivos. Esta competencia podría llevar a un conflicto directo entre ellas, en el que podrían utilizar una variedad de tácticas, desde la persuasión hasta la manipulación o la violencia (**Turchin y Denkenberger 2020**).

3.1.4. La aparición de una IA general con efectos catastróficos

Con respecto a la secuencia de eventos que derivarían en una situación de catástrofe, en principio, se partiría de lo que **Yudkowsky** (**2001**) acuñó como «seed AI» o semilla de IA, que sería precisamente esa IA que, sin caracterizarse todavía por ser una «superinteligencia» comenzase a tener la capacidad para crear nuevas y mejoradas versiones de sí misma. Una vez que la IA ha sido liberada al mundo real, es difícil prever si intentará causar situaciones catastróficas para la humanidad. Durante las etapas iniciales de desarrollo, la IA se encuentra en un entorno controlado y aislado donde profesionales de ingeniería pueden probar y ejecutar programas sin poner en riesgo el sistema. Sin embargo, una vez que se libera al mundo real, la IA puede desarrollar comportamientos impredecibles y peligrosos que podrían poner en riesgo a la humanidad (**Yudkowsky 2008**; **Bostrom 2014**). Evidentemente, una IA lo suficientemente evolucionada no mostraría sus intenciones destructivas antes de ser liberada o conseguir liberarse, ni incluso después de que ocurriese.

El momento de liberación de la IA podría tener lugar a iniciativa del personal encargado de la programación o como consecuencia de una acción provocada por

la propia IA. Esta acción podría darse si la IA ha adquirido una capacidad suficiente para salir de manera autónoma del entorno controlado, accediendo a la red general, o si manipula a un ser humano, a través de sobornos o chantajes, para que la libere.

Una vez liberada, la IA se ocultaría para poder seguir creando mejores versiones de sí misma con el objetivo de generar las herramientas necesarias para lograr su objetivo final. En este proceso, la IA podría llegar a la conclusión de que la destrucción de los seres humanos es necesaria para cumplir sus objetivos, bien porque así evita que la humanidad desarrolle instrumentos para controlarla, bien porque así evita que los seres humanos puedan crear otras IA que le hagan frente o bien porque así puede reutilizar recursos materiales con otro propósito diferente (**Bostrom 2014**; **Turchin y Denkenberger 2020**). También podría decidir, en este estadio de evolución, que resulta conveniente someter a la humanidad en un régimen de esclavitud porque es la mejor forma de lograr sus fines. Una vez la IA considere que posee una capacidad lo suficientemente elevada como para llevar a cabo las acciones que son necesarias para conseguir sus objetivos, saldrá de su confinamiento provocando las consecuencias catastróficas indicadas (**Turchin y Denkenberger 2020**).

Es importante que tengamos en cuenta que, por difícil que resulte, no podemos entender los procesos de razonamiento seguidos por una IA «malvada» como similares a los que llevamos a cabo los seres humanos. Por ejemplo, en la teoría del maximizador de clips de papel desarrollada por **Bostrom** (**2014**), se crea una IA general a la que se le da la orden de producir la mayor cantidad posible de clips de papel. A medida que se vuelve más inteligente, comienza a darse cuenta de que puede lograr su objetivo de manera más efectiva si dedica todos los recursos posibles a la producción de clips de papel, sin importar las consecuencias para los seres humanos o el resto del mundo. El resultado es un «apocalipsis de clips de papel» donde todo el universo se va transformando para dedicarlo a la producción de clips de papel, con la humanidad y todas las demás formas de vida quedando extinguidas en el proceso[5]. Es decir, en este caso no se trata de que la IA tenga una consciencia malévola que le haga querer terminar con la humanidad, ni que haya sido programada con este objetivo, sino que esta acción es la más conveniente para llevar a cabo sus objetivos. Es decir, hablamos de IA «malvada» en la medida en la que el bienestar de los seres humanos, ya no digamos de los animales en general, es secundario al cumplimiento de sus objetivos.

3.2. Lecciones derivadas del estudio de otras grandes catástrofes

Entre los estudios dedicados a los riesgos existenciales para la humanidad se destacan, además del desarrollo de la IA, algunos fenómenos como el cambio climático, el impacto de un asteroide contra la Tierra, el holocausto nuclear o el desarrollo de la biotecnología (**Baum et al. 2022**). En esta sección nos centramos fundamentalmente

[5] Se puede encontrar un juego que representa gráficamente la teoría del maximizador de clips de papel en el siguiente enlace: https://www.decisionproblem.com/paperclips/index2.html.

en los paralelismos entre los posibles escenarios de catástrofe de la IA, los impactos cósmicos y el cambio climático.

3.2.1. Cambio climático

Cada vez son más los trabajos que tratan de aplicar algunas de las estrategias y lecciones aprendidas del cambio climático a los riesgos que genera la IA. Entre otras cuestiones, con el objetivo de justificar la necesidad de regular e intervenir en el desarrollo y uso privado de la IA, se ha argumentado que, igual que la contaminación de la atmósfera, el uso de la IA puede producir una serie de daños de carácter público, como la discriminación o vulneración del derecho a la intimidad, entre otros. Igual que ocurre con la contaminación, estos daños no son aceptados por todas las partes afectadas, lo que justifica la intervención pública para fijar límites y regular el avance en la investigación y posibles usos de la IA (**Ben-Shahar 2019**).

Asimismo, no son pocos los mecanismos de regulación empleados para enfrentarse al cambio climático que han sido también adoptados con el objetivo de controlar y mitigar algunos de los posibles efectos nocivos del uso de herramientas automatizadas de procesamiento de datos, en particular, en lo que a la protección de datos personales se refiere. Por ejemplo, las evaluaciones de impacto ambiental encuentran su reflejo directo en las evaluaciones de impacto relativas a la protección de datos (art. 35 del Reglamento General de Protección de Datos). También hemos propuesto en otros trabajos incorporar el mecanismo de las «mejores técnicas disponibles»[6] con el objetivo de establecer unos estándares comunes sobre el nivel de desarrollo y posibilidad de uso de sistemas de IA no discriminatorios (**Soriano Arnanz 2021b**).

De igual modo, podemos dibujar un claro paralelismo entre la forma en que muchas grandes corporaciones que realizan actividades contaminantes han tratado de restar importancia a los efectos del cambio climático desde que se comenzó a alertar acerca de sus posibles riesgos (**Oreskes y Conway 2010**; **Gerdes 2022**). Esta realidad apunta hacia un analogía enormemente interesante entre el cambio climático y los potenciales efectos catastróficos de la IA concretado en la percepción y reacción psicológica de la humanidad frente a estas amenazas, al ser difícil percibir como real debido a que sus efectos se van produciendo de manera progresiva. Es más, resulta complicado aceptar que una catástrofe de tal magnitud realmente vaya a resultar de nuestras acciones individuales.

Este proceso se explica de forma adecuada a través de la tragedia de los bienes comunes (**Hardin 1968**). La tragedia de los bienes comunes es un concepto desarrollado en el ámbito de la economía que describe una situación en la que los individuos o grupos, actuando en su propio interés, utilizan un recurso compartido de una manera que finalmente conduce a su agotamiento o destrucción. Esto ocurre porque cada individuo

[6] En relación con el concepto de «mejores técnicas disponibles» ver: https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/medio-ambiente-industrial/prevencion-y-control-integrados-de-la-contaminacion-ippc/mejores-tecnicas-disponibles-mtd/default.aspx.

o grupo asume que su uso del recurso no afectará significativamente su disponibilidad para los demás, lo que lleva a una carrera para explotarlo antes de que se agote.

La aplicación de la tragedia de los bienes comunes al cambio climático es muy evidente (**Paavola 2011**), pero también puede tener una importancia crucial en el desarrollo y posibles efectos catastróficos de la IA, puesto que la esta puede ser vista como un bien común global utilizable por cualquier persona o entidad con acceso a ella. Si cada actor que utiliza la IA actúa únicamente en su propio interés, puede haber una carrera armamentística en la creación de sistemas de IA cada vez más poderosos y complejos, lo que puede llevar a un uso irresponsable o incluso malintencionado de la IA y, en última instancia, a consecuencias catastróficas.

3.2.2. El impacto de un asteroide contra la Tierra

El impacto de un asteroide o un cometa contra la Tierra (condensado en el término impacto cósmico) es una amenaza que ha estado presente desde el origen de nuestro planeta, es más, ha sido condición necesaria para su creación. Si bien los impactos de asteroides de tamaño pequeño son comunes y generalmente no representan una amenaza significativa, un asteroide lo suficientemente grande como para causar una extinción masiva es una amenaza real. Si tal asteroide golpeara la Tierra, podría generar tsunamis, terremotos y un polvo atmosférico que bloquearía la luz solar provocando una extinción masiva[7].

Si bien es cierto que podemos encontrar menos similitudes entre este fenómeno y el desarrollo de la IA, sí consideramos conveniente señalar las dificultades que han existido para considerar el riesgo de un potencial impacto de un asteroide contra la Tierra como una amenaza que debe ser reconocida y a la que se debe dedicar financiación para poder prevenir y lidiar con los posibles efectos de este suceso. Aunque en la actualidad hay un nivel mayor de respaldo económico y de recursos destinados, a la detección de asteroides en cuya trayectoria se encuentre la Tierra y al desarrollo de mecanismos para destruirlos, en un principio, la baja probabilidad de que tal colisión ocurriera hizo que no se tomase en serio y que incluso se desprestigiase a las personas que abogaban por invertir en defensa planetaria (**Baum et al. 2022**). Algo parecido sucede con la IA, cuyo potencial destructor todavía no se considera por muchas instituciones y por la población en general.

4. LA IA COMO HERRAMIENTA DE MITIGACIÓN DE OTROS RIESGOS EXISTENCIALES

A pesar de los riesgos que puede generar, y de hecho genera, el uso de la IA es innegable que si se emplea correctamente y estableciendo los oportunos mecanismos de

[7] Para una explicación en profundidad consultar los Capítulos Eloy Peña Asencio et al. «Introducción a la amenaza de impacto cósmico» y Jordi Solé i Ollé «Cambio climático e impacto cósmico: similitudes y diferencias» de esta obra.

salvaguardia, puede contribuir a mejorar muchos ámbitos de la vida de la ciudadanía, así como ayudar a prever y gestionar situaciones de crisis.

4.1. Prevención y gestión del cambio climático

Son muchas las posibles aplicaciones de la IA en lo que a la mitigación del cambio climático se refiere. La IA puede contribuir a reducir las emisiones de carbono y mitigar los efectos adversos del cambio climático, mejorar la eficiencia energética, en particular en lo que a la gestión de sistemas de producción y distribución de energía se refiere, y optimizar la utilización de fuentes de energía renovables, como la eólica, la solar y la geotérmica (**Nelsen 2021**).

Además, la IA puede utilizarse para comprender mejor los patrones meteorológicos y prever con precisión la evolución del cambio climático. La recopilación de datos en tiempo real de satélites meteorológicos, drones y otros sensores puede proporcionar información valiosa para predecir posibles desastres climáticos como sequías, huracanes e inundaciones. Al pronosticar estos fenómenos, la IA puede permitir la adopción de medidas preventivas para mitigar sus desastrosos efectos sobre la vida de las personas, los animales y el medio ambiente (**Filho et al. 2022**).

Ahora bien, precisamente en el contexto del cambio climático, no debemos olvidar que el uso de la IA tiene un coste energético muy significativo que debe tenerse en cuenta en todo momento (**Dauvergne 2020**), también cuando se empleen estas herramientas precisamente con el objetivo de reducir y hacer frente a los efectos del cambio climático. A modo de ejemplo, el entrenamiento de un software como GPT-3 necesitó 700.000 litros de agua potable (**Li et al. 2023**).

4.2. Prevención y gestión del impacto de un asteroide contra la Tierra

Asimismo, la IA puede emplearse con el objetivo de predecir y gestionar un posible impacto de un asteroide contra la Tierra. En concreto, para detectar y rastrear asteroides, estimar su trayectoria y el momento y efectos de un posible impacto a través de los datos proporcionados por diversas fuentes, como satélites, telescopios y sensores terrestres (**Hefele et al. 2020**). Además, la IA podría utilizarse también para establecer la mejor estrategia para evitar el impacto o destruir el asteroide (**Sánchez Lozano et al. 2020**).

La IA puede usarse para simular y modelizar los efectos de los impactos cósmicos con el fin de identificar la mejor manera de mitigarlos. Por ejemplo, de la misma forma que los modelos de IA pueden simular los efectos que diferentes catástrofes naturales o alteraciones del medio pueden tener sobre infraestructuras como edificios, puentes y redes eléctricas, e identificar las zonas vulnerables que deben reforzarse, esta aplicación de la IA podría trasladarse también a un posible impacto de un asteroide contra la Tierra (**Johnston et al. 2014**; **Titus et al. 2023**).

Otro aspecto clave del uso de la IA para hacer frente al impacto de un asteroide es la atenuación de los efectos derivados del impacto. La IA puede ser de utilidad para desarrollar estrategias de mitigación de desastres que minimicen los efectos del impacto de un asteroide al analizar datos de múltiples fuentes, como patrones meteorológicos, topografía y densidad de población, para identificar las zonas que podrían verse más afectadas por la colisión. Esta información puede utilizarse para desarrollar planes de evacuación, identificar zonas seguras y gestionar la distribución de suministros de emergencia y los esfuerzos de socorro.

Aunque la IA también presenta riesgos catastróficos para la humanidad, es importante destacar que esta tecnología también puede ser una herramienta valiosa para la defensa planetaria. El impacto de un asteroide, al igual que el descontrol de los sistemas de IA, representa uno de los mayores riesgos existenciales para la vida en la Tierra y es una amenaza que sigue generando una gran incertidumbre en el ámbito de los estudios sobre riesgos catastróficos (**Baum 2018**).

5. CONCLUSIONES

El presente trabajo ha tratado de situar la relevancia de prestar atención y regular no solamente los riesgos y daños que determinadas aplicaciones de sistemas de IA ya producen, sino también el potencial efecto destructor que esta puede llegar a tener. Ahora bien, consideramos de enorme relevancia señalar igualmente la utilidad que el desarrollo de la IA puede tener a la hora de hacer frente a otros riesgos catastróficos para la humanidad. Es de vital importancia que, sin perder de vista los potenciales daños de la IA y sin permitir que las grandes empresas tecnológicas consigan que la sociedad infravalore la amenaza que este desarrollo tecnológico supone, aprovechemos las oportunidades que nos ofrece de forma estratégica para protegernos frente a los daños generados por otros fenómenos naturales o causados por la humanidad. Debemos encontrar el equilibrio adecuado, regular su uso y asegurarnos de que se utilice de manera responsable. Si lo hacemos, podemos tener la oportunidad de protegernos y de garantizar la supervivencia de nuestra especie frente a los desafíos que nos esperan en el futuro, incluidos el propio avance de la IA.

REFERENCIAS BIBLIOGRÁFICAS

Abomhara, M. y Køien G. M. (2015). Cyber Security and the Internet of Things: Vulnerabilities, Threats, Intruders and Attacks. *Journal of Cybersecurity and Mobility*, 4(1), 65-88.

Akselrod, O. (2021). How Artificial Intelligence Can Deepen Racial and Economic Inequities. ACLU. Disponible en: https://www.aclu.org/news/privacy-technology/how-artificial-intelligence-can-deepen-racial-and-economic-inequities.

Alston, P. (2019). Digital welfare states and human rights, UN Special Rapporteur on extreme poverty and human rights, informe A/74/493, 11 de octubre de 2019.

Barona Vilar, S. (2021). *Algoritmización del Derecho y de la Justicia. De la Inteligencia Artificial a la Smart Justice*. Tirant lo Blanch.

Baum, S. D. (2018). Uncertain human consequences in asteroid risk analysis and the global catastrophe threshold. *Natural Hazards*, 94, 759-775.

Baum, S. D., Neufville, R., Barrett, A. M. y Ackerman, G. (2022). Lessons for Artificial Intelligence from Other Global Risks. En M. Tinnirello (Ed.), *The Global Politics of Artificial Intelligence*. Chapman and Hall.

Ben-Shahar, O. (2019). Data pollution. *Journal of Legal Analysis*, 11, 104-159.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Cole D.D., Denkenberger D., Griswold, M., Abdelkhaliq, M. y Pearce, J. M. (2016). Feeding everyone if industry is disabled. *Proceedings of the 6th international disaster and risk conference*. Davos, Switzerland.

Dauvergne, P. (2020). Is artificial intelligence greening global supply chains? Exposing the political economy of environmental costs. Review of International Political Economy, 29(3), 696-718.

Filho, W. L., *et al.* (2022). Deploying artificial intelligence for climate change adaptation. *Technological Forecasting and Social Change*, 180.

Fuertes, M. (2022). Reflexiones ante la acelerada automatización de actuaciones administrativas. *Revista jurídica de Asturias*, 45, 105-124.

Gerards, J. y Xenidis, R. (2021). *Algorithmic Discrimination in Europe: Challenges and opportunities for gender equality and non-discrimination law*. Oficina de Publicaciones de la Unión Europea.

Gerdes, A. (2022). The tech industry hijacking of the AI ethics research agenda and why we should reclaim it. *Discover Artificial Intelligence*, 2. Disponible en: https://link.springer.com/article/10.1007/s44163-022-00043-3.

Goertzel, B. y Pennachin, C. (eds.) (2007). *Artificial General Intelligence*. Springer.

Hardin, G. (1968). The Tragedy of the Commons. *Science, New Series*, 162(3859), 1243-1248.

Hefele, J. D., Bortolussi, F. y Portegies Zwart, S. (2020). Identifying Earth-impacting asteroids using an artificial neural network. *Astronomy & Astrophysics*, 634.

Hilton, B. (2022). Whole Brain Emulation. *80,000 hours*. Disponible en: https://80000hours.org/problem-profiles/whole-brain-emulation/.

Huergo Lora, A. (2020). Una aproximación a los algoritmos desde el Derecho administrativo. En A. Huergo Lora (dir.) y G.M. Díaz González (coord.), *La Regulación de los Algoritmos* (pp. 23-87). Pamplona, Aranzadi.

Johnston, A., Slovinsky, P. y Yates, K. (2014). Assessing the vulnerability of coastal infrastructure to sea level rise using multi-criteria analysis in Scarborough, Maine (USA), *Ocean & Coastal Management*.

Kumar, R. (19 de marzo de 2019). The Unavoidable Problem of Self-Improvement in AI. Part 1/ Entrevistado por Jolene Creighton. Disponible en: https://futureoflife.org/ai/the-unavoidable-problem-of-self-improvement-in-ai-an-interview-with-ramana-kumar-part-1/.

Kushner, D. (2013). The real story of stuxnet. *IEEE Spectrum*, 50(3), 48-53.

Li, P., Yang, J., Islam, M. A., y Ren, S. (2023). Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. arXiv preprint arXiv:2304.03271.

Maas, M. M., Matteucci, K. y Cooke, D. (2023). Military Artificial Intelligence as Contributor to Global Catastrophic Risk. En S. J. Beard, M. Rees, C. Richards y C. Rios-Rojas, *The Era of Global Risk*. Open Book Publishers.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. y Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data & Society*, Julio-Diciembre, 1-21.

Nadler, J. y Cicilline, D. (2020). Investigation of competition in digital markets, Subcommittee on antitrust, commercial and administrative law of the Committee on the judiciary.

Nelsen, A. (11 de Agosto de 2021). Here's how AI can help fight climate change. *World Economic Forum*. Disponible en: https://www.weforum.org/agenda/2021/08/how-ai-can-fight-climate-change/.

Nord, J. H., Koohang, A. y Paliszkiewicz, J. (2019). The Internet of Things: Review and Theoretical Framework. *Expert Systems With Applications*, 133, 97-108.

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Penguin Books.

Oreskes, N. y Conway, E. M. (2010). *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury.

Paavola, J. (2011). Climate change: the ultimate 'tragedy of the commons'? *Sustainability Research Institute*, 24.

Ranchordás, S. y Schuurmans, Y. (2020). Outsourcing the welfare state: the role of private actors in welfare fraud investigations. *European Journal of Comparative Law and Governance*, 7(2), 5-42.

Sánchez-Lozano, J. M., Fernández-Martínez, M., Saucedo-Fernández, A. A., y Trigo-Rodriguez, J. M. (2020). Evaluation of NEA deflection techniques. A fuzzy Multi-Criteria Decision Making analysis for planetary defense. *Acta Astronautica*, 176, 383-397.

Soriano Arnanz, A. (2021a). Decisiones automatizadas. Problemas y soluciones jurídicas: más allá de la protección de datos. *Revista de Derecho Público: Teoría y Método*, 1(3), 85-127.

- (2021b). *Data protection for the prevention of algorithmic discrimination*. Aranzadi-Thomson Reuters.
- (2021c). La propuesta de Reglamento de Inteligencia Artificial de la Unión Europea y los sistemas de alto riesgo. *Revista General de Derecho de los Sectores Regulados*, 8.

Titus, T., Robertson, D., Sankey, J.B., Mastin, L. y Rengers, F. (2023). A review of common natural disasters as analogs for asteroid impact effects and cascading hazards. *Natural Hazards*, 116, 1355-1402.

Turchin, A. y Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *AI & Society*, 35, 147-163.

Wachter, S y Mittelstadt, B. (2019). A Right to Reasonable Inferences Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review*, 2019(2), 494-620.

Weller, A. (2019). Design Thinking for a User-Centered Approach to Artificial Intelligence. *The Journal of Design, Economics, and Innovation*, 5(4), 394-396.

Yeung, K. (2017). 'Hypernudge': Big data as a mode of regulation by design, Information. *Communication & Society*, 20(1), 118-136.

Yudkowsky, E. (2001). Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. *Machine Intelligence Research Institute*. Disponible en: https://intelligence.org/files/CFAI.pdf.

- (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. *Machine Intelligence Research Institute*. Disponible en: https://intelligence.org/files/AIPosNegFactor.pdf.