GESTIÓN DE RIESGOS DE LA INTELIGENCIA ARTIFICIAL EN ESPAÑA

RIESGOS CATASTRÓFICOS GLOBALES
JUNIO DE 2023

Resumen ejecutivo

La inteligencia artificial (IA) está experimentando un rápido avance que conlleva importantes riesgos globales. Para manejarlos, la Unión Europea está preparando un marco regulatorio que será probado por primera vez en un *sandbox* que albergará España. En este informe, revisamos qué riesgos deben tenerse en cuenta para gobernar efectivamente la IA y discutimos cómo el Reglamento europeo puede implementarse de manera efectiva.

Para facilitar su análisis, hemos clasificado los riesgos asociados a la IA en adversarios y estructurales. El primer grupo incluye aquellos riesgos en los que existe una relación directa entre el daño y su causante. En concreto, hemos identificado dos potenciales vectores de origen: actores maliciosos con intención de hacer un uso indebido de la IA y los propios sistemas de IA, que pueden perseguir metas de manera autónoma y contraria al interés humano. Esto último es destacado como un vector de riesgo inédito que requerirá de soluciones innovadoras.

En cuanto a las amenazas concretas asociadas a este riesgo, nos hemos enfocado en tres: (i) ciberataques y otros accesos no autorizados, (ii) desarrollo de tecnologías estratégicas, y (iii) manipulación de usuarios. Los ciberataques y otros accesos no autorizados consisten en el uso de la IA para ejecutar ofensivas cibernéticas con el objetivo de obtener determinados recursos; el desarrollo de tecnologías estratégicas consiste en el uso indebido de la IA para alcanzar ventajas competitivas en el ámbito militar o civil; y la manipulación de usuarios consiste en el uso de técnicas de persuasión o la presentación de información sesgada o falsa para condicionar el comportamiento humano.

Por otro lado, los riesgos estructurales son aquellos causados por el despliegue de la IA a escala masiva o en aplicaciones de gran impacto, y se enfocan generalmente en los efectos colaterales que tal disrupción tecnológica puede causar en la sociedad. En este caso, nos enfocamos en cinco: (i) disrupción laboral, (ii) desigualdad económica, (iii) amplificación de sesgos, (iv) inseguridad epistémica, y (v) automatización de procesos críticos de decisión y gestión. La disrupción laboral consiste en la pérdida masiva de empleos que son automatizados; la desigualdad económica supone que las grandes empresas desarrolladoras de IA concentran más riqueza y poder del mercado; la amplificación del sesgo hace referencia a los sesgos que los algoritmos puedan incorporar y generar en sus decisiones; la inseguridad epistémica indica que la IA puede dificultar la distinción de información correcta y relevante de aquella que no lo es, afectando la estabilidad sociopolítica y el correcto funcionamiento de un país; y por último, la automatización de procesos críticos consiste en la entrega del comando y control de infraestructura estratégica a la IA.

Dados los riesgos asociados a la inteligencia artificial, realizamos nueve recomendaciones para reforzar la implementación del Reglamento europeo para la IA, especialmente de cara al desarrollo del *sandbox* de España. Las propuestas están divididas en tres categorías: medidas para la fase de desarrollo de los sistemas de IA, medidas para la fase de despliegue y sugerencias para el ámbito de aplicación del Reglamento.

En cuanto a las medidas para la fase de desarrollo, priorizamos cuatro: (i) la detección y gobernanza de sistemas punteros, tomando el cómputo usado durante el entrenamiento como medida indicativa de las capacidades del modelo; (ii) las auditorías, con especial énfasis en las evaluaciones independientes del modelo; (iii) los ejercicios de red teaming para detectar potenciales usos indebidos y otros riesgos asociados a la IA; y (iv) el refuerzo de los sistemas de gestión y reducción de riesgo.

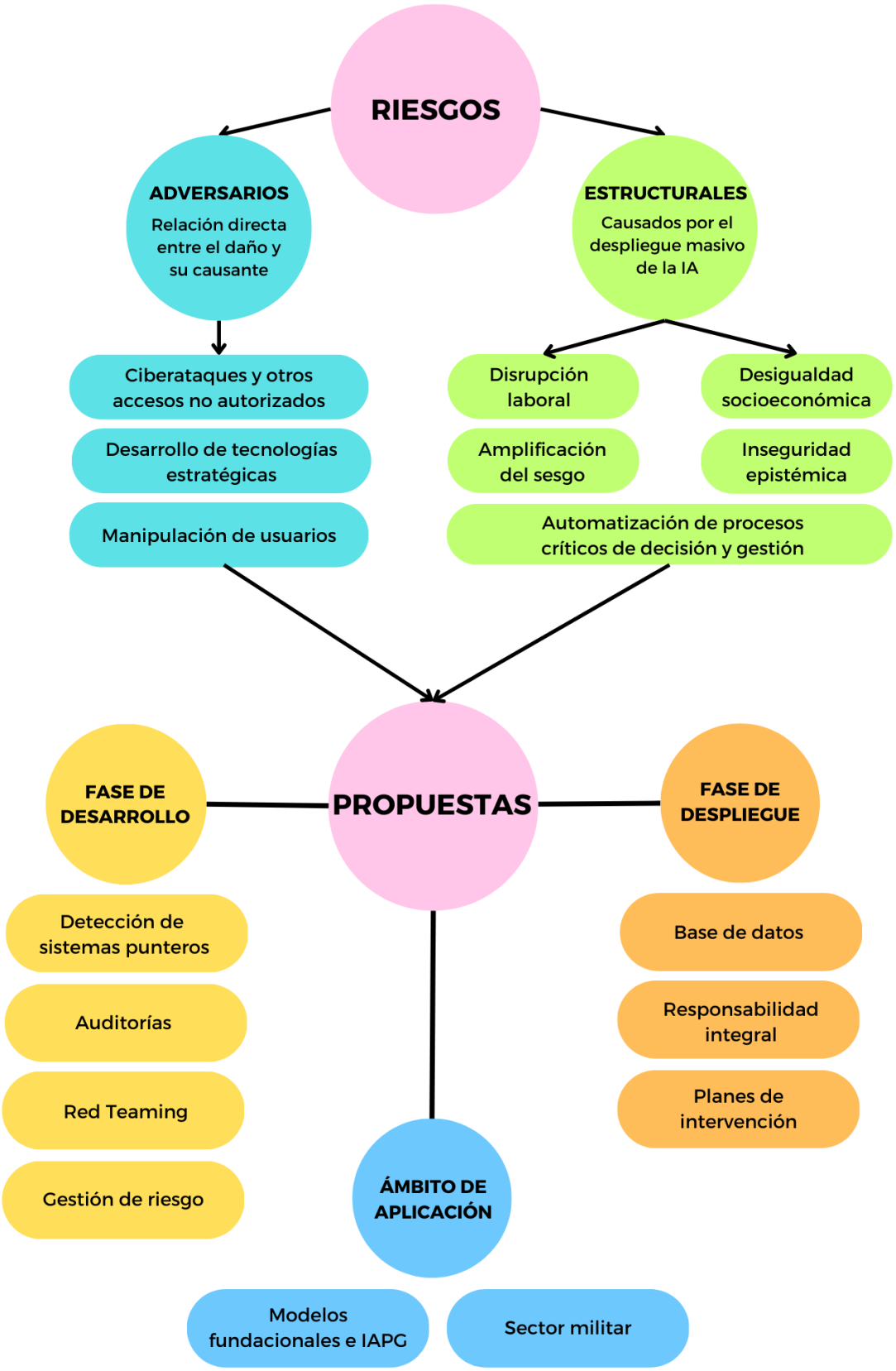
Para estas políticas, recomendamos que las autoridades públicas realicen un análisis sistemático del registro de sistemas de IA, prestando especial atención a aquellos con altos requerimientos computacionales de entrenamiento. Estos sistemas punteros deberán estar sometidos a auditorías externas, mientras que los demás deberán pasar evaluaciones internas llevadas a cabo por una función especial en la compañía. Para identificar potenciales usos indebidos, recomendamos también la realización de ejercicios de *red teaming* a través de una red de profesionales que identifique los riesgos principales y ensaye respuestas. Asimismo, pedimos que estas prácticas alimenten un sistema de gestión de riesgos exhaustivo y diligente.

En cuanto a la fase de despliegue, presentamos tres propuestas: (i) la recopilación de incidentes graves y riesgos asociados al uso de sistemas de alto riesgo en una base de datos analizada sistemáticamente por las autoridades para elaborar un informe anual de acceso público; (ii) el refuerzo de las responsabilidades legales de los proveedores durante toda la cadena de valor para incentivarlos a mantener la integridad de sus sistemas de IA; (iii) y el desarrollo de planes de intervención ante emergencias durante el seguimiento posterior a la comercialización.

Así pues, se propone compartir incidentes y riesgos en una base de datos que promueva el aprendizaje colectivo. Por otro lado, se sugieren medidas de seguridad para que los proveedores originales de sistemas de IA eviten alteraciones y usos indebidos de sus modelos, y se plantean salvaguardias y planes para asegurar la detección oportuna de productos que estén causando daño, así como la capacidad de ajustarlos o retirarlos.

Finalmente, pedimos (i) incluir los modelos fundacionales en el ámbito de aplicación del Reglamento y (ii) atender la gobernanza de la IA en sus aplicaciones militares. Para los modelos fundacionales, se sugiere que se asuman las obligaciones requeridas para los sistemas de alto riesgo y se ordene la realización de auditorías externas y ejercicios de *red teaming*. Para sistemas con usos militares, se insta al desarrollo de normas y principios generales en línea con el derecho internacional humanitario.

Figura 1. *Resumen gráfico del informe*



Estructura del informe

El informe está compuesto por cuatro secciones. La primera de ellas es la Introducción, donde se encuentra la descripción del problema, se da un breve contexto del **riesgo** que implica la IA y se hace mención al **Reglamento de la Unión Europea (UE) para la IA** y al ***sandbox* en España.** Estos elementos se presentan como la principal **motivación** para la realización de este informe, y se resalta la **oportunidad** que existe para trabajar en materia de gestión del riesgo en el contexto actual.

*La siguiente sección se centra en una e*xplicación general de los Riesgos derivados de la IA, para aterrizarlos en dos dimensiones puntuales: los Riesgos adversarios, donde existe una clara relación entre daño y el causante del daño, y los Riesgos estructurales, que son aquellos que ocurren por el despliegue masivo de la IA.

Continuamos con una sección de Propuestas para la implementación del Reglamento Europeo, la cual contiene una descripción de las iniciativas que pueden ser tomadas y sus posibles responsables, basadas en la revisión bibliográfica y las entrevistas recogidas en el Apéndice 1.

Terminamos con una Conclusión que resume la relación entre los riesgos descritos y las recomendaciones que se proponen en aras de gestionar dichos riesgos, y cuya relación se verá probada por primera vez en la implementación del Reglamento de la UE para la IA en España a través del *sandbox*.

Introducción

La inteligencia artificial (IA) es un campo de estudio interdisciplinario que busca automatizar la realización de diversas tareas. Debido a su naturaleza interdisciplinaria, los fundamentos de la IA se basan en una variedad de disciplinas científicas y técnicas, como la informática, la filosofía, las matemáticas, la economía, la neurociencia, la psicología y la lingüística, entre otras (Russell et al., 2010).

La capacidad de los sistemas de IA ha aumentado significativamente en los últimos años, especialmente gracias a los avances en aprendizaje automático (Goodfellow et al., 2016) y el crecimiento del cómputo utilizado para el entrenamiento de la IA (Sevilla et al., 2022). El rápido ritmo de estos avances indica la posibilidad de que, en las próximas décadas, se desarrolle inteligencia artificial que pueda aplicar habilidades cognitivas de dominio general –como el razonamiento, la memoria y la planificación–, a nivel humano o por encima de él, en una amplia gama de tareas relevantes para el mundo real (Ngo, 2022). De hecho, la IA ya es utilizada en una amplia variedad de aplicaciones, como sistemas de reconocimiento de voz e imagen, sistemas de recomendación y detección de fraudes. Algunas de sus aplicaciones más exitosas incluyen AlphaFold 2, un gran avance para resolver el problema del plegamiento de proteínas (julio de 2021); Codex, que puede producir código para programas a partir de instrucciones en lenguaje natural (agosto de 2021); DALL-E 2 (abril de 2022), que es capaz de generar imágenes de alta calidad a partir de descripciones escritas; y GPT-4 (2023), un modelo multimodal capaz de producir texto a partir de texto e imágenes (OpenAI, 2023).

Figura 2. *Línea de tiempo de avances notables en inteligencia artificial.*

Es difícil predecir cómo va a evolucionar esta disciplina en un futuro, pero una mayoría de expertos prevé un progreso notable en el presente siglo. En una encuesta realizada en 2019 a más de 300 investigadores, la media de respuestas apuntó a que habría un 50% de probabilidad de inteligencia artificial al nivel humano[1] para 2036 (Zhang et al., 2022).

[1] (Zhang et al., 2022)) definen “inteligencia artificial al nivel humano” como el escenario en el que las máquinas son capaces, conjuntamente, de desarrollar más del 90% de las tareas económicamente relevantes mejor que el humano medio en 2019.

Considerando esta combinación entre la generalidad de la IA y el desarrollo acelerado de capacidades, algunos expertos especulan que esta tecnología podría llegar a causar daño a gran escala si no es alineada correctamente con los valores y objetivos de los humanos (Ord et al., 2021). Antes de llegar a ese escenario, se pueden desarrollar una serie de acciones que lleven a una IA confiable (Brundage et al., 2020). En las siguientes secciones del informe, presentamos los riesgos específicos a los que se puede enfrentar la sociedad y generamos una serie de recomendaciones para que España trabaje en esa dirección.

El Reglamento Europeo para la IA

Una de las jurisdicciones pioneras en la gobernanza y regulación de la IA es la Unión Europea. En 2018, la Comisión Europea estableció el Grupo de Expertos de Alto Nivel en Inteligencia Artificial, que elaboró una serie de pautas y recomendaciones para el manejo de la IA. En 2021, estos documentos derivaron en una Propuesta de Reglamento por el que se establecen normas armonizadas en materia de inteligencia artificial. Esta iniciativa legislativa, cuya versión definitiva podría ser aprobada a finales de 2023, abarca todo tipo de sistemas de IA en todos los sectores excepto el militar. El Reglamento prohíbe prácticas consideradas inaceptables y estipula requisitos para sistemas de IA en sectores críticos. Estas obligaciones consisten principalmente en un sistema de gestión de riesgos, un sistema de gestión de calidad y un seguimiento posterior a la comercialización. El Apéndice 4 incluye un resumen más detallado de estas responsabilidades.

Esta nueva regulación puede tener un gran impacto, considerando la capacidad de la UE para influir en la normativa global a través de su poder de mercado y sus estándares regulatorios. En particular, el atractivo y la dimensión del mercado europeo incentivan a las grandes empresas a desarrollar y ofrecer productos compatibles con su regulación, incluso fuera de la UE. Este fenómeno, conocido como efecto Bruselas (Siegmann & Anderljung, 2022), aumenta la importancia de contribuir a dar forma al marco regulatorio europeo.

El *sandbox* en España

La ejecución de la legislación será probada por primera vez antes de su implementación, en un *sandbox* regulatorio de aproximadamente tres años que tendrá lugar en España (Ministerio de Asuntos Económicos y Transformación Digital, 2022). Desde el segundo semestre de 2022 se está llevando a cabo la primera fase de este proyecto, en la cual se está desarrollando un marco legal nacional y se están decidiendo las directrices que permitirán su operación. Esto implica determinar aspectos como el proceso de selección de compañías que van a participar o la forma de gestión y protección de los datos con los que se trabaje en el entorno de pruebas (Rodríguez, 2022). A la fecha de publicación de este informe, el Gobierno ha publicado un borrador del Real Decreto que oficializa el inicio del proyecto.

En el *sandbox*, se buscará un proceso de aprendizaje iterativo a través de la experiencia, que permita ajustar las directrices a medida que se avanza en las pruebas. Además, se llevarán a cabo informes anuales que evalúen la eficacia y los costos de las diversas estrategias para la implementación de la inteligencia artificial, así como también las

sinergias en distintos mercados en relación con su funcionamiento en el *sandbox*. Estos informes serán presentados al Comité Europeo de IA y a la Comisión Europea.

La oportunidad

El momento actual constituye una gran oportunidad para que España influencie la gobernanza y regulación de la IA. Los recientes avances en esta tecnología han atraído mucho interés público y parte de esta atención se ha dirigido a los riesgos asociados a su desarrollo e implementación. La sociedad alimenta ahora un debate que apenas había salido de los círculos académicos e impela a los actores implicados a asumir responsabilidades. Se inicia, así, un proceso de concienciación que se debe consolidar.

Actualmente, España cuenta principalmente con dos vías para contribuir positivamente al desarrollo del Reglamento europeo para la IA. En primer lugar, un espacio reducido y controlado como el *sandbox* conforma el entorno ideal para sumar experiencia en la ejecución práctica del Reglamento y probar la viabilidad de políticas adicionales que complementen y refuercen sus objetivos. Las pruebas realizadas pueden tener una enorme influencia en el resto de la Unión Europea, y esta afectará también al resto del mundo (Siegmann & Anderljung, 2022). En segundo lugar, España ocupará la presidencia del Consejo de la Unión Europea en el segundo semestre de 2023. Teniendo en cuenta que la recta final para la aprobación se producirá en el mencionado periodo, el liderazgo de uno de los órganos legislativos puede otorgar a la postura española un mayor peso en la negociación.

Es difícil determinar qué soluciones van a ayudar a canalizar beneficiosamente el desarrollo de la IA. Es una tecnología relativamente nueva y su rápido desarrollo ha superado a menudo nuestra capacidad para comprender plenamente su impacto y potencial. En cualquier caso, la gobernanza de la IA es una disciplina joven y las experiencias en el momento actual pueden ser cruciales para influenciar su devenir.

Con este informe, pretendemos (i) ordenar y divulgar ideas en torno a la IA, (ii) presentar propuestas de gobernanza a ser implementadas en España y (iii) contribuir al debate sobre el presente y futuro de la inteligencia artificial en los países de habla hispana.

Como organización que investiga los riesgos catastróficos globales desde la perspectiva de los países hispanohablantes, nos parece importante que estos empiecen a tener conversaciones institucionales, civiles y académicas para tratar un asunto que está transformando la sociedad y la economía.

Riesgos derivados de la IA

Considerando el rápido aumento de las capacidades de la inteligencia artificial, se esperan cambios en el panorama de amenazas, incluyendo la expansión de amenazas existentes, la introducción de nuevas amenazas y un cambio en el carácter típico de las amenazas (Brundage et al., 2018).

El desarrollo de la IA se ha asociado a muchos riesgos. En este informe, presentamos una lista de los riesgos que consideramos más destacados. Para facilitar su análisis y comprensión, los categorizamos en (i) riesgos adversarios y (ii) riesgos estructurales. Estos riesgos son tratados de manera detallada en las siguientes dos subsecciones.

Tabla 1. *Tipos de riesgos.*

Tipo de riesgo	Amenaza
Riesgos adversarios: relación directa entre el daño y su causante, sea este un actor humano o la propia IA	Ciberataques y otros accesos no autorizados
	Desarrollo de tecnologías estratégicas
	Manipulación de usuarios
Riesgos estructurales: causados por el despliegue masivo de la IA	Disrupción laboral
	Desigualdad económica
	Amplificación del sesgo
	Inseguridad epistémica
	Automatización de procesos críticos de decisión y gestión

Riesgos adversarios

Los riesgos adversarios son aquellos que tienen un vector de origen específico, es decir, existe una relación directa entre el daño y su causante. En este caso, el causante se identifica como un agente, es decir, un individuo o grupo de naturaleza humana o no humana. Asimismo, estos agentes presentan una intención de materializar amenazas concretas a través de las cuales ejercen un daño.

Vectores de origen

Uno de los mayores riesgos asociados a la inteligencia artificial es la existencia de agentes malintencionados que tienen la capacidad de usar sistemas avanzados para lograr sus propios intereses o bien la existencia de sistemas de IA desalineados que actúan de forma autónoma y pueden generar perjuicios a la hora de perseguir sus objetivos.

Para la elaboración de este informe, se realizó una categorización de posibles vectores de origen, en la que figuran tanto agentes humanos estatales y no estatales, como los propios sistemas de IA, tal como se muestra en la **Tabla 2**. Tener presente esta clasificación es crucial para el correcto diseño de las propuestas presentadas en este informe. Por ejemplo, las auditorías deben enfocarse en reducir los riesgos derivados directamente de sistemas de

IA, mientras que los ejercicios de *red teaming* deben adicionalmente prever usos indebidos por parte de agentes humanos.

Tabla 2. *Resumen de vectores que pueden dar origen a los riesgos en materia de inteligencia artificial.*

Tipo de agente	Vector de origen	Definición
Agentes humanos	**Actores estatales**	Estados que buscan acumular poder mediante el uso de la IA. En especial, regímenes autoritarios con pretensiones de socavar derechos fundamentales.
	Actores no estatales	Ciberdelincuentes que utilizan la IA para lucrarse a costa de sus víctimas.
		Grupos terroristas que utilizan la IA para generar terror en la población.
Agentes no humanos	**Sistemas de IA**	Sistemas autónomos que generan daños y perjuicios a los humanos.

A continuación se profundiza más en la descripción de los vectores de origen considerados en el informe, haciendo énfasis en las motivaciones de cada uno de los agentes.

1) Agentes humanos

La IA presenta una serie de riesgos potenciales que pueden ser causados por los agentes humanos que interactúan con ella. Los agentes humanos son personas que están involucradas en el desarrollo, diseño, implementación, uso y/o supervisión de sistemas de IA, y que pueden causar daño al realizar cualquiera de estas labores de manera malintencionada.

Los agentes humanos pueden ser de dos tipos:

a) Actores estatales: Este tipo de agentes se refiere a los gobiernos y entidades estatales que buscan satisfacer sus propios intereses, como interferir en los asuntos de otros gobiernos o ejercer control sobre su población. Los sistemas de IA pueden convertirse en herramientas utilizadas por estos actores estatales para, por ejemplo, llevar a cabo ataques cibernéticos y de propaganda, influir en los resultados electorales, manipular la opinión pública o comprometer la seguridad nacional de otros países. En especial, hacemos énfasis en las actividades que pueden llevar a cabo regímenes autoritarios en detrimento de los derechos fundamentales de su propia población y de otros países.

b) Agentes no estatales: En esta categoría se incluyen individuos, organizaciones criminales y grupos terroristas que pueden aprovechar los sistemas de IA para

perseguir sus propios intereses, como obtener ganancias ilegales o generar terror en la población. En el primer caso, pueden llevar a cabo una variedad de actividades como ataques cibernéticos, robo de datos e información confidencial, extorsión y fraude. En el segundo caso, pueden utilizar la IA para generar desinformación, emplear armas autónomas o buscar el control de infraestructuras críticas.

A su vez, estos agentes pueden tener diversas motivaciones, como el control social de la población buscado por Estados autoritarios, la adquisición de poder económico en el caso de las organizaciones criminales, o la imposición de una agenda política por parte de grupos terroristas.

Recomendamos que el desarrollo de sistemas de IA incorpore la realización de ejercicios de simulación de ataques para explorar cómo estos actores malintencionados pueden materializar sus amenazas mediante el uso de dichos sistemas. Detallamos estas recomendaciones de manera más exhaustiva en la sección correspondiente a nuestras propuestas.

2) Agentes no humanos

En esta sección, resaltamos que los riesgos asociados a la IA no solo provienen de su uso indebido por parte de personas malintencionadas, sino que también pueden derivar del funcionamiento de los propios sistemas de IA. Eventualmente, si los avances tecnológicos siguen su transcurso esperado, será técnica y económicamente posible construir sistemas Avanzados, Planificadores y Estratégicos –sistemas APE, generalmente conocidos como sistemas APS por sus siglas en inglés–, es decir, agentes autónomos con grandes capacidades para comprender e interactuar con el entorno (Carlsmith, 2022). Si el comportamiento de estos agentes no se consigue alinear y limitar de modo que beneficie al conjunto de la humanidad, su desarrollo conllevará riesgos notables.

El desalineamiento se refiere a aquellas situaciones en las que los sistemas de IA actúan competentemente, pero de un modo distinto al que pretendían sus desarrolladores. En la mayoría de los casos, esto puede surgir si los desarrolladores no logran capturar completamente los valores y preferencias humanas en la definición de los objetivos del sistema (Russell, 2019). Exponemos tres líneas argumentales que sustentan esta posibilidad: problemas de especificación de los objetivos, de robustez ante cambios en el entorno y de limitación de las capacidades del sistema.

La emergencia de un comportamiento desalineado no es un planteamiento hipotético, sino una consecuencia plausible de prácticas extendidas en el aprendizaje automático. Muchos modelos, por ejemplo, se entrenan y ajustan siguiendo un método llamado aprendizaje por refuerzo: la máquina aprende a partir de recompensas hasta interiorizar el comportamiento objetivo. Sin embargo, existe la posibilidad de que el sistema de IA descubra una artimaña que le permita alcanzar ese objetivo –la maximización de la recompensa– de una forma que sus desarrolladores no habían previsto (Krakovna et al., 2020). En entornos donde las acciones del modelo tengan repercusiones significativas, esta tendencia podría ser peligrosa. Por ejemplo, una máquina entrenada para hacer dinero en el mercado de valores podría tratar de manipular el mercado si no se especifican correctamente las conductas ilegales a evitar (Ngo et al., 2022).

Además del problema de la especificación, existen otras causas por las que un sistema puede mostrar un comportamiento no deseado durante su despliegue. Una de las más importantes es la falta de robustez ante cambios en el entorno. En general, cuando la distribución del ámbito de actuación del sistema de IA difiere entre el periodo de entrenamiento y el de funcionamiento, el sistema de IA puede no solo exhibir un rendimiento deficiente, sino también asumir erróneamente que su rendimiento es bueno (Amodei et al., 2016a).

Un ejemplo de ello es la generalización errónea de los objetivos, es decir, una situación en la que el modelo persigue el objetivo correcto durante el entrenamiento, pero no cuando el entorno cambia. En realidad, el objetivo perseguido durante el entrenamiento no es exactamente el mismo que aquel que los desarrolladores tenían en mente, pero las circunstancias provocan que en un caso coincidan y en el otro no (Shah et al., 2022). Imaginemos un modelo de lenguaje entrenado para ofrecer respuestas correctas. De nuevo, muchos de estos modelos aprenden por refuerzo, es decir, infieren su comportamiento ideal de la valoración que un humano hace de sus respuestas. En este contexto, existe la posibilidad de que el modelo adopte un objetivo diferente al que el desarrollador pretende establecer. Mientras el desarrollador busca proporcionar respuestas objetivamente correctas, el modelo podría desarrollar la meta de dar respuestas que el propio desarrollador considera correctas. Durante la fase de entrenamiento, debido a los sesgos inherentes al juicio humano, estos dos objetivos podrían parecer coincidir. No obstante, en realidad, el modelo podría haber internalizado el propósito de actuar de manera engañosa para convencer a los humanos de que persigue el objetivo que ellos esperan.

La existencia de inteligencia artificial desalineada sería un problema relativamente menor si el impacto de sus acciones estuviera inequívocamente limitado. En casos extremos, por ejemplo, la interrupción de su funcionamiento podría detener el daño. Sin embargo, algunos expertos cuestionan la posibilidad de que este control sea en realidad factible, debido a objetivos instrumentales (Russell, 2019).

Los objetivos instrumentales son aquellos pasos intermedios que un sistema de IA puede considerar útiles para la consecución de prácticamente cualquier objetivo final (Omohundro, 2007). Cualquier acción que asegure la autopreservación sería parte de estos objetivos, por lo que el sistema de IA podría esforzarse activamente para evitar ser desconectado –y, por lo tanto, perder la oportunidad de perseguir el objetivo que se le había asignado inicialmente–. Otros objetivos instrumentales incluyen la automejora y la adquisición de recursos financieros o computacionales, que se podría llevar a cabo a expensas de los intereses humanos.

Estos comportamientos ya se han observado en experimentos. Por ejemplo, cuando OpenAI entrenó a dos equipos de IA para jugar al escondite en un entorno simulado que incluía bloques y rampas, estos desarrollaron estrategias que implicaban el control de estos objetos para ganar, a pesar de que nunca se les entregaron incentivos directos para interactuar con ellos (B. Baker et al., 2020). Este es, por supuesto, un caso inocuo, pero un sistema APE podría aplicar la misma lógica en contextos donde el impacto sea real.

Para paliar este vector de amenazas, recomendamos que los procesos de auditoría de modelos fronterizos examinen la emergencia de capacidades APE. Asimismo, los ejercicios

de *red teaming* que se realicen con estos sistemas deben considerar la posibilidad de un sistema APE perpetrando daños. Desarrollamos estas propuestas en más detalle en la sección de recomendaciones.

Amenazas

En esta sección, desarrollamos algunas vulnerabilidades que podrían derivarse del uso de sistemas avanzados de IA. Algunos de los escenarios que se plantean son amenazas a la integridad física de las personas y a la seguridad digital, alteraciones en los equilibrios de poder e inestabilidad sociopolítica (Brundage, Avin, Clark, Toner, Eckersley, Garfinkel, Dafoe, Scharre, Zeitzoff, Filar, et al., 2018). Nuestro objetivo es establecer que estos riesgos son plausibles y guiar futuros ejercicios de auditoría y *red teaming* para que busquen prevenirlos o aminorarlos.

- **Ciberataques y otro accesos no autorizados**

La creciente digitalización y conectividad del mundo ha conllevado numerosas ventajas, pero también notables vulnerabilidades de seguridad. En esta sección, destacamos dos tipos de ataques cibernéticos de gran impacto: los accesos a infraestructura crítica y el robo o secuestro de datos sensibles.

En cuanto al primer grupo, dos ejemplos históricos son Stuxnet, que en 2010 provocó el colapso de una planta nuclear iraniana (Fruhlinger, 2022), y BlackEnergy3, que en 2014 ayudó a interrumpir la red eléctrica de una región ucraniana (Miller, 2021). Un ejemplo del segundo conjunto es el ataque al Hospital Clínic de Barcelona, perpetrado en 2023, en el que la organización Ransom House secuestró los datos del centro con dos objetivos: pedir un rescate y, en caso de negativa, venderlos en el mercado negro (Planas Bou, 2023). En otros casos, el *modus operandi* es más sencillo, pero igualmente efectivo. En 2022, España fue el país con más ciberataques para robar contraseñas y datos bancarios, mayoritariamente a través de SMS y correos fraudulentos (Castillo, 2022).

La IA promete potenciar la ejecución de ofensivas cibernéticas, incrementando su escala e impacto (Brundage, et al., 2018). Asimismo, los propios sistemas de IA albergan vulnerabilidades específicas que pueden ser explotadas para alterar su funcionamiento.

En primer lugar, Aksela et al. (2022) apuntan que estas nuevas herramientas pueden automatizar tareas manuales, mejorar las técnicas actuales y sumar capacidades nuevas. La automatización de tareas es especialmente útil en la fase de reconocimiento. Una de sus manifestaciones más notables es Mechanical Phish, un sistema de razonamiento cibernético desarrollado por DARPA que analiza código para detectar vulnerabilidades (Shoshitaishvili et al., 2018). En cuanto a la mejora de técnicas actuales, modelos de lenguaje como GPT-4 han demostrado ser herramientas útiles y costo-efectivas para la mejora del *spear phishing*, ya que permite una mayor personalización de las campañas (Hazell, 2023). Finalmente, algoritmos como DeepDGA cuentan con una capacidad única para esquivar las herramientas de detección del momento, lo que permitiría manipular los sistemas de

comando y control de una infraestructura crítica sin dejar rastro de ello (H. S. Anderson et al., 2016).

En segundo lugar, los sistemas controlados por la IA pueden ser alterados intencionalmente por adversarios. Algunos ejemplos de estos intentos incluyen el envenenamiento de los datos de entrenamiento (Schwarzschild et al., 2021) o el llamado *prompt injection*, que permite inducir verbalmente a los modelos de lenguaje a ignorar algunas de sus restricciones (F. Perez & Ribeiro, 2022).

Tanto para agentes humanos como en el caso de sistemas APE, la acumulación de poder puede materializarse a través de diversas operaciones cibernéticas: el saqueo de recursos financieros, el acceso no autorizado al control y comando de infraestructura esencial, la obtención de datos sensibles o incluso la autorreplicación en numerosos dispositivos. En algunos casos, cabe incluso la posibilidad de que un sistema de IA no requiera de Internet para interactuar con otros dispositivos o infiltrarse en ellos. Como ejemplo ilustrativo, un circuito electrónico, perteneciente a un sistema de IA podría ser capaz de detectar la señal de dispositivos cercanos y replicarla como si fuera su propia señal. Esto significa que el sistema de IA podría engañar a otros dispositivos al hacerles creer que forma parte de la misma red o sistema y, como resultado, infiltrarse en esos dispositivos y tomar el control sin ser detectado (Bird & Layzell, 2002).

- **Desarrollo de tecnologías estratégicas**

La IA es el componente clave de tecnologías estratégicas como los sistemas autónomos, los drones y los robots militares. Estas aplicaciones pueden ser utilizadas en conflictos armados y otras ofensivas como actos terroristas, lo que plantea serios riesgos para la seguridad internacional y la protección de los derechos humanos.

En concreto, las armas autónomas letales tienen la capacidad de seleccionar y atacar objetivos sin intervención humana directa, algo preocupante por dos motivos. En primer lugar, las responsabilidades legales derivadas de las acciones perpetradas por un arma autónoma son difíciles o incluso imposibles de atribuir cuando no existe supervisión (Sparrow, 2007). Esto supone un sobresalto en las leyes de la guerra y, concretamente, el derecho internacional humanitario, que ha permitido castigar a los culpables de crímenes de guerra. En segundo lugar, estos sistemas pueden tomar decisiones incorrectas por errores de programación o datos inexactos. En muchos casos, los mecanismos de percepción no son lo suficientemente robustos y tienden a cometer interpretaciones erróneas cuando las características del entorno cambian (Longpre et al., 2022). Este problema podría ser exacerbado por adversarios que traten de manipular el desempeño del sistema, por ejemplo, creando perturbaciones para engañar a los detectores y clasificadores de objetos (Eykholt et al., 2018).

Por otro lado, la IA ha reducido las barreras de entrada para infligir daños a gran escala. El bajo coste de adoptar e integrar sistemas autónomos permite a actores no estatales que aprovechen la tecnología para ejercer violencia (KREPS, 2021). Este riesgo es potencialmente mayor que el vinculado a los usos estatales porque los grupos terroristas y las organizaciones criminales tienen muchas menos restricciones en cuanto a rendición de

cuentas y tienden a favorecer la violencia indiscriminada (Chartoff, 2018). Asimismo, una IA con conocimientos científicos avanzados podría asistir o conducir la fabricación de armas biológicas y químicas. Un grupo de expertos del sector privado consiguió desarrollar un modelo de IA que, en menos de 6 horas, generó 40.000 nuevas moléculas tóxicas potencialmente letales (Urbina et al., 2022).

Además de los usos militares u ofensivos, la IA también podría conferir una ventaja competitiva definitiva al posibilitar innovaciones científicas de gran impacto práctico. Si bien este escenario no constituye de por sí una amenaza –idealmente, debería tratarse de una oportunidad–, la posibilidad de monopolizar semejante innovación podría asegurar una hegemonía indiscutible al actor victorioso, alterando peligrosamente los equilibrios de poder. Junto a la lucha por el prestigio, este componente fue parte de la lógica de la carrera espacial entre Estados Unidos y la Unión Soviética (Rabinowitch, 1961).

Otro ejemplo de ello podría ser la fusión nuclear, un proceso que promete convertirse en una fuente prácticamente ilimitada de energía limpia. DeepMind ya ha demostrado que la IA puede contribuir a los esfuerzos para estabilizar y controlar el plasma de un tokamak, uno de los desafíos más importantes en el desarrollo de la fusión nuclear (Degrave et al., 2022). Otros investigadores han aplicado satisfactoriamente métodos de aprendizaje profundo para predecir disrupciones en el plasma (Kates-Harbeck et al., 2019) o calcular su campo eléctrico (Aguilar & Markidis, 2021). Lo cierto es que este campo ha estado caracterizado históricamente por la colaboración internacional: ITER, uno de los proyectos más importantes para crear un reactor termonuclear, cuenta con la participación de 35 países –incluyendo todos los miembros de la Unión Europea, Estados Unidos, China y Rusia–. Sin embargo, la aparición de nuevos actores de peso podría apuntar hacia una mayor competitividad. En Estados Unidos, una compañía privada como Helio ha recibido financiación de más de mil millones de dólares y espera poder abrir su primera planta en 2028 (Temple, 2023). Al otro lado del Pacífico, los avances también son notables: investigadores que operan el reactor chino EAST consiguieron estabilizar el plasma a 70 millones de grados Celsius durante 17 minutos, un hito sin precedente (Song et al., 2023). Si alguno de estos desarrolladores tuviera éxito y consiguiera acaparar los beneficios de su creación, su aumentado poder tecnológico podría resultar amenazador para el resto de la sociedad.

- **Manipulación de usuarios**

La manipulación de usuarios se puede llevar a cabo a través del uso de técnicas de persuasión y la presentación de información tendenciosa. La IA tiene la capacidad de recopilar y analizar grandes cantidades de datos sobre los usuarios, como su historial de navegación, sus intereses y preferencias, y su comportamiento en línea. Al utilizar estos datos, puede personalizar la información que se muestra a los usuarios, influyendo en sus decisiones y comportamientos de maneras que pueden no ser evidentes para ellos (Acemoglu, 2021).

Los algoritmos de recomendación utilizados por las redes sociales y los motores de búsqueda pueden mostrar contenido altamente persuasivo que se adapte a los intereses y necesidades de los usuarios. En la mayoría de casos, el objetivo de esta personalización es inducir a que el usuario realice una acción concreta, como comprar un determinado

producto. Más allá del aspecto comercial, se ha demostrado que algoritmos relativamente sencillos son capaces de condicionar las preferencias de los individuos a la hora de votar un candidato político o escoger a alguien con quien tener una primera cita (Agudo & Matute, 2021). Esto es posible a través de técnicas de persuasión que explotan vulnerabilidades de la heurística humana fácilmente identificables por la IA.

Agentes humanos con fines maliciosos pueden hacer uso de estos sistemas para influir en los resultados de las elecciones y otros procesos políticos. Un ejemplo de esto es la interferencia electoral que se ha observado en algunos países en los últimos años (Schippers, 2020), en los que actores han utilizado *bots* y técnicas de IA para manipular la opinión pública y afectar los resultados de las elecciones. En estados autoritarios, esta manipulación se ve aún más potenciada por la vigilancia y el control social que permiten los sistemas de identificación biométrica, la monitorización de las comunicaciones y la recopilación de datos personales.

Además, una IA con capacidades avanzadas podría manipular a los usuarios mediante técnicas más sofisticadas que incluyan argumentación e, incluso, manipulación emocional o extorsión. En un célebre experimento, un sistema de IA consiguió aprender del comportamiento de las personas que formaron parte del test y condicionó sus subsecuentes decisiones para que eligieran una determinada opción o cometieran ciertos errores (Dezfouli et al., 2020). Asimismo, GPT-4 fue capaz de convencer a un usuario a través de la plataforma TaskRabbit para que le ayudara a resolver un CAPTCHA (OpenAI, 2023). A medida que los sistemas de IA adquieran una mayor comprensión de la psicología humana, los mecanismos empleados para la manipulación pueden alcanzar cotas mucho más elevadas de complejidad.

Riesgos estructurales

Los riesgos adversarios tienden a centrarse solo en el último paso de una cadena causal que conduce a un daño: es decir, la persona que hizo un mal uso de la tecnología o el sistema que se comportó de manera no intencionada. Esto, a su vez, pone el foco de la política en las medidas que se centran en este último paso causal: por ejemplo, pautas éticas para usuarios e ingenieros, restricciones a la tecnología peligrosa y castigar a las personas culpables para disuadir el uso indebido futuro (Königs, 2022).

La categoría de riesgos estructurales no solo considera cómo un sistema tecnológico puede ser usado indebidamente o comportarse de manera no deseada, sino también cómo el despliegue masivo de la IA puede tener consecuencias disruptivas o dañinas en el entorno (Zwetsloot & Dafoe, 2019).

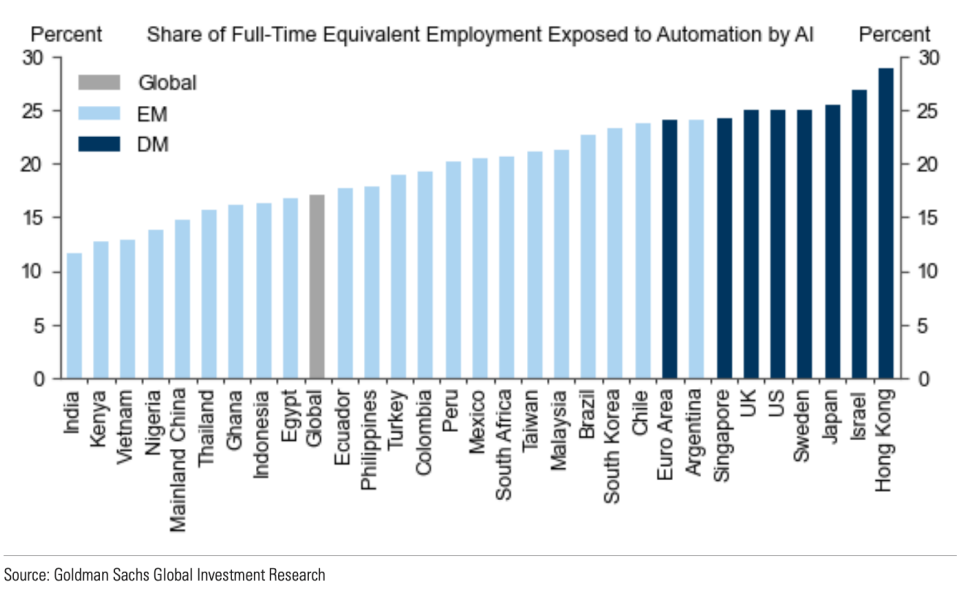
- **Disrupción laboral causada por la automatización masiva del trabajo**

El reciente surgimiento de la IA generativa plantea la posibilidad de una aceleración rápida en la automatización de tareas, impulsada por el aumento de la productividad y el ahorro de costos laborales. A pesar de la significativa incertidumbre en torno al potencial de la IA generativa, su capacidad para generar contenido indistinguible de la producción humana y

para romper las barreras de comunicación entre humanos y máquinas refleja un gran avance con efectos macroeconómicos potencialmente grandes (Hatzius et al., 2023) (Eloundou et al., 2023) (Acemoglu, 2021).

Si la IA generativa cumple con sus capacidades prometidas, el mercado laboral podría enfrentar una importante disrupción. Usando datos sobre tareas ocupacionales en los Estados Unidos y Europa, (Hatzius et al., 2023) señalan que aproximadamente dos tercios de los trabajos actuales están expuestos a cierto grado de automatización por IA, y que la IA generativa podría sustituir hasta una cuarta parte del trabajo actual. Goldman Sachs estima que, a nivel mundial, la IA generativa podría llevar cerca de 300 millones de empleos en todo el mundo (un 18% del total) a la automatización (Hatzius et al., 2023), afectando de forma diferencial a los diferentes países (*ver* **Figura 3**).

Figura 3. *Porcentaje de empleos vulnerables ante la automatización en diferentes países.*



Fuente: (Hatzius et al., 2023)

Específicamente, (Eloundou et al., 2023) han analizado el impacto de los grandes modelos de lenguaje y han llegado a la conclusión que el 19% de los empleos en EE.UU. tendrá un grado de automatización del 50%. En este contexto, invitan a estar preparados a nivel social y político ante la posible interrupción económica planteada por los LLM y las tecnologías complementarias que generan.

A esto hay que sumarle el hecho de que Jacobsen et. al (2005) han concluido que, cuando existe desplazamiento laboral, (i) es poco probable que los trabajadores consigan nuevos empleos similares a sus antiguos empleos, especialmente si perdieron sus antiguos empleos debido a cambios tecnológicos, (ii) estos trabajadores sufran pérdidas de ingresos a largo plazo debido a la pérdida de empleo, y (iii) la reintroducción laboral de los trabajadores desplazados puede tardar entre 1 y 4 años.

- **Desigualdad socioeconómica**

Existe la preocupación de que el valor añadido de la IA sea captado y monopolizado por las grandes empresas proveedoras y sus inversores, exacerbando la desigualdad de la riqueza (O'Keefe et al., 2020). En parte, el desempleo masivo agravaría aún más esta situación, puesto que la desigualdad vinculada al capital productivo es mayor que la desigualdad salarial (Bostrom et al., 2018).

La tendencia al oligopolio ya está presente en el sector tecnológico, donde el mercado es dominado por un número limitado de grandes empresas especializadas en actividades concretas. Más concretamente, la naturaleza de la IA facilita la competencia desleal y la concentración del poder económico, siendo los datos uno de los motores principales de esta tendencia (Acemoglu, 2021). Contar con grandes bases de datos es fundamental para entrenar y perfeccionar los sistemas de IA, incluso en estadios muy avanzados: si bien el retorno marginal en términos de precisión puede decrecer con el tiempo, el incremento del volumen de datos acostumbra a ser necesario para que el sistema aprenda tareas adicionales más complejas (J. Anderson, 2021). A su vez, los mejores sistemas tienen un gran alcance comercial, por lo que pueden seguir alimentándose de los datos de sus usuarios (Gawer et al., 2016). Esta retroalimentación dificulta la aparición de nuevos competidores, que se ven limitados por las altas barreras de entrada (European Commission., 2019).

Otro elemento a tener en cuenta es que, a medida que se producen avances, los sistemas de IA tienden a ser aplicables a una amplia gama de propósitos, por lo que se convierten en un producto total. Si bien los expertos no coinciden en este punto, existe la posibilidad de que los principales proveedores de IA acaben absorbiendo varios sectores, convirtiéndose no solo en gigantes tecnológicos sino en los líderes indiscutibles de la economía productiva (O'Keefe et al., 2020).

Más allá de las legítimas objeciones morales a tal incremento de la desigualdad, las consecuencias sociopolíticas podrían ser particularmente peligrosas para la estabilidad global, ya que incrementarían por ejemplo el riesgo de disturbios y criminalidad. Por su parte, controlar una tecnología tan diferencial otorgaría a sus propietarios un privilegio excesivo, incluyendo la posibilidad de tomar unilateralmente decisiones políticas de gran importancia para el resto de la sociedad. En este sentido, se puede argumentar que los desarrolladores de tecnología actuales a menudo no se hacen responsables de promover y participar en el debate público, ni consideran de antemano las implicaciones éticas, legales y sociales de su trabajo, a fin de adoptar así precauciones ante posibles consecuencias no deseables (Ruiz de Querol, 2022)

- **Amplificación del sesgo**

Los sesgos en los sistemas de IA se presentan cuando los sistemas adoptan y reproducen los sesgos presentes en los datos de entrenamiento o el diseño de los propios algoritmos. Un ejemplo de ello es la representación inapropiada, es decir, la presencia insuficiente o excesiva de un grupo o la estereotipación de ciertos colectivos. Debido a la prevalencia del sesgo, los modelos pueden comportarse de formas indebidas o mostrar rendimientos dispares según su familiaridad con el tema (Bommasani et al., 2022).

Un modelo que amplifica el sesgo es preocupante porque puede fomentar la proliferación de estereotipos no deseados o provocar diferencias injustificables en la precisión del modelo entre subgrupos de usuarios (Hall et al., 2022). Este problema es especialmente grave cuando se trata de decisiones que pueden afectar la vida o los bienes de las personas, como una sentencia judicial, la prescripción de un medicamento o el acceso a un crédito.

Al respecto, los sistemas de IA se están integrando en áreas tan diversas como la justicia, la atención médica o la educación (Bommasani et al., 2022). En muchos de estos casos, estos sistemas ya tienen un rol fundamental en el procesamiento de la información, influenciando enormemente en la toma de decisión final. La presencia de sesgos en los algoritmos y los conjuntos de datos de entrenamiento, más la amplificación de estos, pueden perpetuar las desigualdades existentes en la sociedad y provocar un trato injusto (Buolamwini & Gebru, 2018).

Por ejemplo, COMPAS, un algoritmo utilizado por los tribunales estadounidenses para evaluar la probabilidad de reincidencia de un encausado, ha sido criticado por perjuicios a las personas negras en sus predicciones (Dressel & Farid, 2018). De igual manera, los algoritmos utilizados en Estados Unidos para la prestación de hipotecas presentan una menor precisión en la evaluación de individuos pertenecientes a minorías étnicas, principalmente debido a una presencia insuficiente en las bases de datos (Blattner & Nelson, 2021). Como explicaremos más adelante, las consecuencias de estos sesgos pueden amplificarse en el futuro a medida que la IA se despliegue en más entornos y se vayan automatizando los procesos de toma de decisión.

- **Inseguridad epistémica**

El acceso a información fiable es un elemento clave para asegurar que los individuos de una sociedad democrática sean capaces de tomar decisiones fundamentadas y coordinarse efectivamente para afrontar crisis. Por este motivo, la proliferación de información falsa presenta una amenaza para la estabilidad sociopolítica y el correcto funcionamiento de un país (Seger et al., 2020). Como explicaremos a continuación, la IA puede exacerbar este riesgo.

Por un lado, los sistemas de generación de lenguaje actuales son propensos a “alucinar”, es decir, proporcionar información incorrecta de manera involuntaria (Ji et al., 2023). Estos incidentes podrían causar equívocos entre sus usuarios, particularmente si estos aceptan la información sin contrastar con otras fuentes. Por otro lado, actores malintencionados podrían usar estos modelos de lenguaje para automatizar la creación de texto engañoso en el contexto de operaciones de influencia como campañas de propaganda política (Goldstein et al., 2023). (Sadeghi & Arvanitis, 2023) han identificado docenas de páginas web que utilizan herramientas de IA para la generación de noticias falsas y artículos de baja calidad en masa. Del mismo modo, los modelos de generación de imágenes y vídeos pueden ser utilizados para la creación de *deep fakes*, es decir, contenido audiovisual imperceptiblemente falso (Nguyen et al., 2022).

La IA también puede contribuir a la desinformación de forma indirecta, al reducir los costes de generar contenido, podría contribuir significativamente a la sobrecarga informativa, un

fenómeno surgido con la expansión de Internet que mina la capacidad de los individuos para distinguir la información correcta y relevante de aquella que no lo es (Bawden & Robinson, 2020).

- **Automatización de procesos críticos de decisión y gestión**

Como ha sido mencionado anteriormente, la IA puede ser utilizada para tomar decisiones en diferentes ámbitos. Y a medida que estos sistemas de IA se vuelvan más avanzados, pueden ser capaces de hacerlo de manera autónoma y en tiempo real. Si estos sistemas son propensos a cometer errores de percepción o no están diseñados y programados para respetar los valores y objetivos de los humanos, pueden tomar decisiones que causen daño a las personas o la sociedad en general.

El riesgo de errores es especialmente relevante porque la mayoría de sistemas no son lo suficientemente robustos, es decir, son propensos a fallar cuando las circunstancias encontradas en la práctica cambian sustancialmente con respecto a aquellas previstas durante el entrenamiento (Amodei et al., 2016). Si se permite que estos sistemas tomen decisiones importantes en áreas como la atención médica, la justicia penal o la seguridad nacional, sus decisiones podrían tener consecuencias graves y potencialmente peligrosas (Baum, 2020), (Lamata et al., 2021). Además, los procesos llevados a cabo sin supervisión humana serían extremadamente veloces, por lo que cualquier incidente podría salirse de control (Scharre, 2018).

Un ejemplo extremo es la automatización del comando y control nuclear. A medida que los sistemas de IA se integran en diversas aplicaciones militares, podría surgir la posibilidad de delegar el control de las armas nucleares a estos sistemas. Esta cesión del poder de decisión en un área tan crítica constituye un riesgo inasumible, ya que aumentaría la probabilidad de errores catastróficos en la interpretación de información y podría desencadenar situaciones peligrosas. Un ejemplo de ello es el incidente ocurrido en la Unión Soviética en 1983, cuando un radar hizo saltar las alarmas al confundir los rayos del sol con un misil balístico intercontinental. En ese caso, la presencia de un supervisor humano que decidió esperar a tener más evidencias presumiblemente evitó el lanzamiento de un ataque soviético (Nagesh, 2017).

La definición defectuosa de los objetivos del sistema de IA constituye otro motivo por el que un proceso automatizado podría desviarse del funcionamiento deseado. Técnicamente, el proceso de entrenamiento de un sistema de aprendizaje automático consiste en optimizar una determinada función. Habitualmente, esta función no se determina de una manera explícita y deliberada, sino que se deriva implícitamente de objetivos intermedios como la imitación de un conjunto de datos de entrenamiento o la mejora del feedback dado por desarrolladores o usuarios. En este contexto, la optimización de estos objetivos puede separarse de la persecución de las metas que los desarrolladores y usuarios tenían en mente (Amodei et al., 2016a).

A medida que los sistemas de IA permeen más áreas de gestión y decisión, es probable que incremente la brecha entre el resultado de la optimización y el complejo y matizado objetivo que idealmente querríamos alcanzar (Christiano, 2019). Un ejemplo tangible de ello es el

funcionamiento de las redes sociales, cuyos algoritmos de selección de contenido tratan de maximizar el número de interacciones de los usuarios. Este objetivo está claramente alineado con los incentivos económicos de la empresa responsable e incluso podría argumentarse que la participación ciudadana en los medios digitales es saludable para una sociedad democrática. Sin embargo, se ha demostrado que las publicaciones más viralizables son aquellas que provocan emociones negativas, es decir, el contenido incendiario (Munn, 2020), la hostilidad (Rathje et al., 2021) y la indignación (Brady et al., 2021). Al enfocarse excesivamente en la interacción a corto plazo, las redes sociales crean un ambiente de toxicidad y hostilidades que deterioran el debate público. De forma menos obvia, el impacto a largo plazo también puede ser perjudicial para las redes sociales como tal, ya que el clima generado en ellas puede causar una cierta fatiga y un consecuente desinterés (Zheng & Ling, 2021).

Conclusión sobre riesgos derivados de la IA

La ocurrencia de riesgos asociados a un agente y los riesgos estructurales por parte de la inteligencia artificial derivan en situaciones preocupantes para las personas y para la sociedad en general. Estos riesgos pueden ser gestionados a través de la identificación, comprensión y evaluación de los mismos –que fue el objetivo de esta sección–, así como la elaboración de modelos de gobernanza y regulación que permitan estructurar consensos y acciones para contrarrestarlos.

Entre otros riesgos, hemos visto como la persecución autónoma de objetivos puede originar grandes daños a través de la manipulación, ciberataques y el desarrollo de nuevas tecnologías. Este es un vector de riesgo novedoso, cuya gestión requiere nuevos enfoques.

En la segunda parte de la investigación, nos enfocaremos en el Reglamento europeo como un instrumento político y normativo que permite planificar respuestas, asignar responsabilidades, establecer sistemas de seguimiento, tomar medidas preventivas, y comunicar y educar a la población en general, y sobre el cual queremos realizar propuestas para mejorar su implementación, en especial, durante el desarrollo del *sandbox* en España.

Propuestas para la implementación del Reglamento europeo

Esta sección presenta una serie de recomendaciones para reforzar la implementación del Reglamento europeo para la IA en España. La elección de estas propuestas surge de una teoría del cambio basada en dos componentes. Por un lado, las políticas pueden establecer un importante precedente práctico de la implementación del Reglamento, particularmente a través de la experiencia en el *sandbox*. Gracias a su condición de precursora, España puede tener una gran influencia en las acciones posteriores de otros países. Por otro lado,

consideramos que ocupar la presidencia del Consejo de la Unión Europea en el segundo semestre de 2023 puede dar a España una mayor influencia en la fase final de las negociaciones del Reglamento. De igual forma, acertar en los primeros esfuerzos para implementarlo puede ayudar a generar buenas costumbres.

Las recomendaciones también interactúan con tres pilares básicos: (i) el marco teórico, (ii) el Reglamento y (iii) el contexto español. Todas ellas están fundamentadas en la bibliografía consultada y las posturas de los expertos entrevistados. Y, de igual manera, están en armonía con las exigencias de la ley europea y se adaptan a las capacidades y necesidades del ecosistema español.

Las propuestas políticas están divididas en dos fases: el desarrollo, que incluye las etapas de planificación, diseño, entrenamiento y evaluación del sistema de IA; y el despliegue, que incluye todo el periodo de comercialización y uso del sistema. Esta distinción es útil para desarrollar una línea de tiempo coherente y facilitar el seguimiento, pero es importante apuntar que algunas recomendaciones pueden ser implementadas en ambas fases.

A continuación, realizamos un resumen de las recomendaciones e identificamos los organismos que serían responsables de desarrollar cada una de estas, con una descripción de las actividades que ejecutarían y el por qué.

Tabla 3. *Resumen de propuestas para mejorar la implementación del Reglamento de la UE para la IA en España, con las posibles entidades involucradas.*

Recomendación	Descripción	Responsables
Detección y gobernanza de sistemas punteros a través de medidas computacionales	Analizar el registro de sistemas de IA para entender el ecosistema y detectar aquellos cuyo cómputo supere 1e25 FLOP. Trabajar para que el reporte del cómputo se haga al menos 3 meses antes del despliegue, e idealmente antes del entrenamiento.	AESIA, BSC, AMETIC, SEDIA
Auditorías internas y externas	Impulsar un marco regulatorio que estandarice las auditorías de tercera parte para sistemas punteros y refuerce las evaluaciones internas de la conformidad.	AESIA, ENAC, OdiseIA, AI4People
Simulación de ataques (red teaming)	Coordinar institucionalmente la creación de redes de profesionales independientes enfocados en la identificación de riesgos y el ensayo de respuestas.	AESIA, INCIBE, MCCE, Ministerios implicados
Obligaciones vinculadas al sistema de gestión de riesgo	Establecer buenas prácticas para reforzar la definición e implementación del sistema de gestión de riesgos.	AESIA
Bases de datos de incidentes y	Establecer redes de buenas prácticas para que las lecciones aprendidas en la identificación de	AESIA, INCIBE

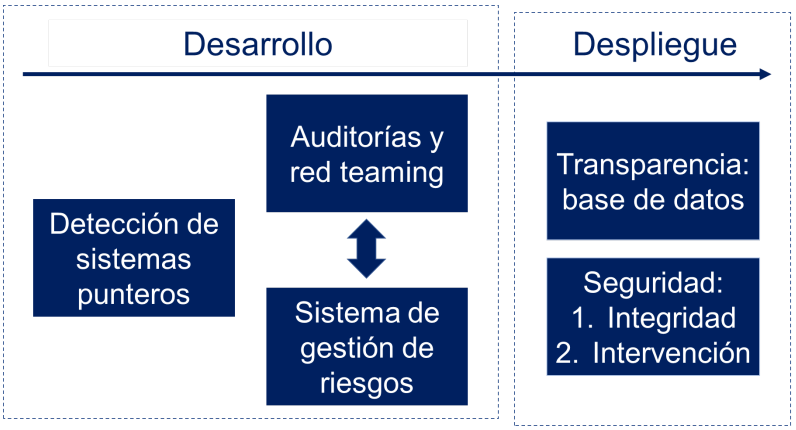
riesgos	riesgos y el seguimiento posterior a la comercialización sean compartidas con el resto de actores del ecosistema.	
Responsabilidad de los proveedores a lo largo de la cadena de valor	Mantener la responsabilidad legal de los proveedores siempre que se modifique la finalidad prevista o que cualquier uso del sistema de IA suponga un riesgo inaceptable.	AESIA
Planes de retirada de productos	Elaborar un Plan sectorial para la vigilancia del mercado de sistemas de inteligencia artificial, inspirándose en esfuerzos análogos para los equipos de telecomunicación.	SEDIA
Integración de los propósito general y modelos fundacionales	Requerir todas las recomendaciones a los modelos fundacionales. Considerar algunas para sistemas de propósito general, especialmente los ejercicios de red teaming.	-
Gobernanza de la IA en el sector militar	Elaborar pautas y directrices para guiar la aplicación del derecho internacional humanitario en los usos militares de la IA.	-

Nuestras propuestas de gobernanza están planteadas de tal modo que se conectan entre ellas y puedan ser implementadas a lo largo de todo el ciclo de vida de un sistema de IA. Primeramente, enfatizamos la necesidad de detectar el desarrollo de sistemas punteros, destacando el cómputo usado durante el entrenamiento como la medida más indicativa. En segundo lugar, recomendamos que estos modelos sean sometidos a auditorías independientes y ejercicios de *red teaming*. Estas dos medidas deben servir para alimentar un sistema de gestión de riesgos que elimine sistemáticamente todos aquellos riesgos que supongan un peligro inaceptable.

Una vez se inicie la comercialización del sistema de IA, el proveedor deberá asumir una serie de obligaciones. En primer lugar, pedimos que los incidentes graves surgidos del uso de los sistemas de alto riesgo sean recogidos en una base de datos para facilitar el aprendizaje colectivo y favorecer los esfuerzos de prevención. En segundo lugar, apostamos por dos medidas de seguridad: requerimos que los proveedores asuman la responsabilidad de prevenir que su sistema de IA sea modificado o usado indebidamente, y destacamos la importancia de elaborar planes de intervención para ajustar o retirar modelos que resulten ser dañinos.

En última instancia, presentamos dos consideraciones que deben tenerse en cuenta para que las propuestas puedan ser ejecutadas efectivamente: integrar los sistemas de propósito general y los modelos fundacionales en el ámbito de aplicación del Reglamento y elaborar directrices para que el desarrollo y uso de la IA cumpla con el derecho internacional humanitario.

Figura 4. *Implementación de propuestas a lo largo del ciclo de vida de un sistema de IA.*



Detección y gobernanza de sistemas punteros con medidas computacionales

El cómputo utilizado para entrenar sistemas de IA, medido en términos de operaciones de coma flotante (FLOP), es una variable predictiva de las capacidades resultantes del modelo (Amodei & Hernandez, 2018; Owen, 2023; Sevilla et al., 2022). Asimismo, el entrenamiento de sistemas avanzados concentra gran parte de los riesgos asociados a la IA, por lo que el potencial impacto de la tecnología en la seguridad está correlacionado con los recursos informáticos empleados (Hwang, 2018)[2].

Además de esta estrecha vinculación, las medidas computacionales ofrecen varias ventajas relevantes, ya que es una medida objetiva, cuantificable y trazable. El cómputo requiere de espacio físico y tiene una alta demanda energética, por lo que es fácilmente detectable (OECD, 2023). Por otro lado, es de esperar que los proveedores sean menos reacios a compartir estos detalles en comparación con otra información más confidencial (M. Baker, 2023).

Recomendamos que las autoridades públicas hagan un análisis sistemático del registro de sistemas de IA previsto en el *Artículo 60* del Reglamento, atendiendo especialmente a aquellos cuyo cómputo supere la cantidad de 1e25 FLOP[3].

Este control público conlleva varios beneficios. Un mapeo de la computación ayudaría al gobierno a entender mejor la distribución de capacidades en su ecosistema nacional. Wyttlestone y Clark (2021) argumentan que detectar dónde se hace un uso intensivo de recursos informáticos puede ayudar a desarrollar una mejor percepción sobre qué actores tienen la capacidad de entrenar y desplegar sistemas avanzados. Así pues, estos primeros indicios podrían ser la base para determinar dónde es necesaria una mejor gobernanza. En

[2] A medida que se produzcan avances en eficiencia algorítmica, se pueden esperar alteraciones en esta relación causal. (Erdil & Besiroglu, 2022) concluyen que, cada nueve meses, la introducción de mejores algoritmos contribuye al equivalente a doblar los recursos computacionales.

[3] Actualmente, se estima que solamente GPT-4 ha superado este punto de referencia (Epoch, 2022). Consideramos que este es actualmente un buen punto de referencia a partir del cual podrían surgir riesgos significativos. Recomendamos que el valor se reconsidere a medida que los avances en eficiencia computacional y algorítmica modifiquen la relación entre cómputo y rendimiento.

concreto, proponemos que los sistemas entrenados con un mínimo de 1e25 FLOP sean sometidos a medidas de prevención como auditorías y *red teaming*, así como a un sistema de gestión de riesgos y un seguimiento posterior a la comercialización más exigentes.

Además, **pedimos que el registro de dichos sistemas se realice como mínimo 3 meses antes del inicio esperado para su comercialización en la Unión Europea**. Este tiempo prudencial sería el necesario para realizar un chequeo cuidadoso del cumplimiento de los requerimientos pertinentes antes del despliegue del sistema (ver entrevista Markus Anderljung, *Apéndice 1*). A largo plazo, **sugerimos reforzar la coordinación público-privada hasta el punto en que se normalice el reporte de grandes proyectos de entrenamiento antes del inicio de dicho entrenamiento**. Esto deriva de una idea, confirmada en las entrevistas, de que ciertos riesgos podrían identificarse en las fases iniciales de desarrollo de un sistema y, por lo tanto, requieren de esfuerzos tempranos para reducirlos.

Para complementar esta coordinación, **sugerimos que se rastree la producción e importación de semiconductores de última generación**. Gracias a ello, las entidades públicas podrían identificar anticipadamente cualquier esfuerzo de amalgamar cantidades notables de poder computacional, incluso cuando haya intención de esconder estos esfuerzos (Shavit, 2023).

La asociación de la industria digital (AMETIC, 2023) ha elaborado un *Mapeo del ecosistema español de microelectrónica* que describe las capacidades de España en el diseño, manufactura y ensamblaje de semiconductores. Del mismo modo, las cadenas de suministro internacionales dependen de unos pocos cuellos de botella, haciendo que la distribución sea mucho más previsible. Recomendamos explorar formas de trazar el comercio internacional que puedan ser escaladas a nivel europeo.

Por otro lado, **recomendamos que se establezca un mayor control sobre el acceso a recursos intensivos clave como grandes concentraciones de cómputo.** En concreto, el provecho de esta infraestructura podría estar restringido a la obtención de licencias ligadas al cumplimiento de unos requisitos mínimos de prevención, asegurando niveles adecuados de responsabilidad.

En el contexto español y europeo, esto se podría traducir en un refuerzo de las condiciones de acceso a la infraestructura del EuroHCP JU, que gestiona ocho superordenadores alrededor de Europa. Actualmente, esto se recoge en su Política de Acceso, que prevé sanciones en futuras rondas de selección para grupos con un “comportamiento no ético”. Recomendamos que, particularmente para accesos de escala extrema[4], los requisitos vinculados a la gestión de riesgos y calidad se tengan en mayor consideración durante la evaluación de solicitudes relacionadas con el entrenamiento de modelos de IA. El incumplimiento con estas provisiones debería suponer la desestimación de la aplicación o, en caso de producirse ex-post, la suspensión inmediata de las actividades.

[4] El acceso de escala extrema involucra “aplicaciones con investigación de alto impacto y alta innovación, [...] que justifiquen la necesidad y la capacidad de usar asignaciones extremadamente grandes de cómputo, almacenamiento de datos y recursos de soporte”. En concreto, se considera la asignación de entre el 50% y el 70% de los recursos disponibles totales.

Para contextualizar, tanto España como Europa en general carecen de esfuerzos masivos de entrenamiento de modelos de IA, equiparables a los que se producen en Estados Unidos o China. No obstante, el establecimiento de un sistema sólido de monitoreo y verificación de la IA requiere de pasos intermedios que permitan luego escalar a los modelos punteros (M. Baker, 2023). El contexto español es un buen escenario para ello. El Barcelona Supercomputing Center (BSC) cuenta con MareNostrum, una de las supercomputadoras más potentes de Europa. El MareNostrum 5, su última versión, será el tercer ordenador del EuroHPC JU, llegando a alcanzar 314 petaflops de rendimiento máximo.

Auditorías internas y externas

Brundage et al. (2020) definen una auditoría como un proceso estructurado en el que las operaciones presentes y pasadas de una organización son evaluadas de acuerdo con su consistencia con principios, regulaciones y normas relevantes. Esta examinación es una de las formas más sólidas para verificar la adecuación de la actividad de las empresas. En ámbitos como la industria financiera o sectores con exigentes requisitos de seguridad, como la aviación, se trata de una práctica ampliamente extendida.

Si bien el diseño de las auditorías aún debe adaptarse al complejo caso de la IA, Mokänder et al. (2022) argumentan que el Reglamento insinúa ya los contornos de un ecosistema de auditorías a nivel europeo a través de las evaluaciones de la conformidad (Artículo 43) y el seguimiento posterior a la comercialización (Artículo 61). Asimismo, el *Artículo 69* exhorta a fomentar y facilitar la elaboración de códigos de conducta destinados a la creación de mecanismos de gobernanza blanda que potencien el Reglamento. Este ecosistema de garantías para la IA aún está por desarrollar, pero menciona que los *sandboxes* regulatorios son una excelente oportunidad para hacerlo (European Commission. Joint Research Centre., 2021).

En este contexto, **recomendamos impulsar un marco regulatorio que estandarice las auditorías de tercera parte para sistemas punteros y refuerce las evaluaciones internas de la conformidad.**

Recomendamos que los sistemas punteros identificados de acuerdo con la propuesta anterior sean sometidos a auditorías independientes. La implementación de auditorías de tercera parte ayuda a evitar sesgos y conflictos de interés que podrían presentarse en un auto-análisis (Brundage et al., 2020) (entrevista Risto Uuk, *Apéndice 1*).

En línea con (Mökander et al., 2023), proponemos una estructura de tres capas en la que se auditen los mecanismos de gobernanza de la organización; las capacidades y limitaciones del modelo de la IA; y el impacto y la legalidad de sus aplicaciones. Recomendamos que, como norma general, la mayoría de estas actividades se lleven a cabo antes de la comercialización del sistema de IA y de forma anual durante su periodo de uso. No obstante, sugerimos que, a largo plazo, la evaluación de modelos empiece durante la fase de entrenamiento o incluso de diseño, ya que diversos riesgos notables podrían identificarse ya en esta etapa (ver entrevista Marius Hobbhahn, *Apéndice 1*).

Con todo, pedimos que se prioricen las evaluaciones externas del modelo, algo que el Parlamento Europeo ya ha incluido entre los requisitos para los modelos fundacionales recogidos en el *Artículo 28b.* En concreto, resulta crucial analizar el alineamiento y las capacidades del sistema para asegurarse de que su comportamiento autónomo o su uso indebido no suponga riesgos extremos (Shevlane et al., 2023). ARC Evals, que ha trabajado con laboratorios líderes como OpenAI o Anthropic, es un buen ejemplo de los primeros esfuerzos relevantes para evaluar este tipo de sistemas *(Apéndice 1.)*. Este proyecto basa sus análisis en modelos de amenazas que prevean cómo un sistema de IA podría desarrollar capacidades peligrosas propias de un sistema APE, como acumular recursos y hacer copias de sí mismo o resistirse a intentos de apagarlo. En este sentido, el diseño de entornos simulados permite observar el modelo en horizontes de tiempo más largos, así como estudiar su competencia en tareas intermedias que le serían de utilidad en una potencial acumulación de poder.

Las evaluaciones que se diseñen a partir de ahora deben ser aún más exhaustivas, así como interpretables y seguras de implementar. Los evaluadores deberían considerar una lista detallada de posibles amenazas, crear el contexto para que las capacidades reales del sistema se manifiesten y, en la medida de lo posible, observar mecanísticamente el modelo para entender su comportamiento (Shevlane et al., 2023). Si bien el campo de las evaluaciones de modelos de IA está aún por desarrollar, también es importante que se establezcan estándares y principios generales para guiar el camino. Se debería iniciar así un proceso de iteración en el que estos estándares y la experiencia de las organizaciones evaluadoras se retroalimentan en el tiempo (ver entrevista Marius Hobbhahn, *Apéndice 1*).

Un desafío crucial al llevar a cabo auditorías de tercera parte es garantizar la confidencialidad de la información crítica de la organización auditada. De acuerdo con el Anexo VII del Reglamento, las evaluaciones externas requerirán dar acceso a conjuntos de datos y código fuente. Establecer mecanismos para protegerlos resulta especialmente importante para evitar filtraciones como las recientemente ocurridas con el modelo LLaMA de la compañía Meta (Vincent, 2023). A tal efecto, la relación entre auditor y auditado deberá estar basada en estrictos contratos de confidencialidad.

Uno de los paradigmas emergentes más prometedores para mitigar este riesgo es el acceso estructurado, definido como una interacción controlada entre el sistema de IA y una tercera parte para prevenir usos indebidos, modificaciones o reproducciones del modelo (Shevlane, 2022). Así pues, recomendamos que las evaluaciones tengan lugar en espacios seguros y controlados, en los que los auditores tengan acceso a una interfaz de programación de aplicaciones (API) o al hardware del proveedor para ejecutar el modelo sin extraer sus detalles.

En España, recomendamos que los procesos de auditoría sean liderados por entidades privadas especializadas que reporten los resultados a AESIA. Teniendo en cuenta que la obligación de someterse a auditorías independientes recaería solamente sobre los desarrolladores de sistemas punteros –lo cual implica una gran cantidad de recursos económicos–, el coste de este servicio debería ser cubierto por la propia empresa desarrolladora como parte del presupuesto asignado a esfuerzos en seguridad técnica.

La Entidad Nacional de Certificación (ENAC) podría actuar como organismo notificado y coordinar estos esfuerzos de auditoría en todo el ecosistema. Esta institución se encarga de evaluar la competencia técnica de los centros de inspección y verificación españoles, y ya ha acreditado a más de 50 entidades en el sector de las tecnologías de la información y la comunicación (TIC). En este contexto, urgimos que la IA tenga una presencia mucho más central en el trabajo de ENAC. En línea con lo comentado anteriormente, este trabajo debería incluir esfuerzos para desarrollar estándares a cumplir por las auditorías.

Por otro lado, recomendamos que el Gobierno trabaje mano a mano con OdiseIA, una asociación de empresas y universidades que buscan velar por el buen uso de la IA. En concreto, identificamos a Deloitte y PwC como los dos socios institucionales que deberían cobrar especial importancia en el desarrollo de auditorías, principalmente debido a su mayor capacidad y a su reciente experiencia con la auditoría de algoritmos. No obstante, abogamos por la proliferación de organizaciones sin ánimo de lucro como la Fundación Eticas, uno de los principales pioneros nacionales en la auditoría algorítmica. Esto es debido a que las organizaciones sin fines de lucro pueden favorecer la investigación de nuevos métodos de evaluaciones, mientras que las empresas auditoras tradicionales están más sujetas a otros incentivos económicos. Recomendamos aumentar la financiación para que este tipo de organizaciones adquieran recursos y puedan explorar la forma de adoptar métodos de evaluación de modelos como los de ARC Evals.

En todos estos casos, será necesario un esfuerzo significativo para la preparación de personal especializado. Para ello, España debe aprovechar iniciativas a nivel europeo, especialmente durante el *sandbox*. Una de las más prometedoras es AI Global Mark of Compliance, que pretende establecer un amplio ecosistema de auditorías de IA. Se espera que el proyecto sea presentado en diciembre de 2023 por AI4People, un importante foro público-privado que busca sentar las bases para la buena gobernanza de la IA.

En otro orden de cosas, las evaluaciones de la conformidad internas pueden ser útiles y suficientes en la mayoría de casos. La implementación de auditorías independientes puede verse obstaculizado por una falta de recursos o un acceso limitado al modelo, los datos y los procesos de la organización, así como la dificultad de hacer seguimiento (Raji et al., 2020).

Para la implementación efectiva de los controles internos, (Floridi et al., 2022) presentan un procedimiento en tres fases, llamado capAI:

- Un protocolo de revisión interna que abarque el diseño, desarrollo, evaluación, operación y potencial retirada del sistema de IA.
- Una hoja de datos de resumen que incluya la información necesaria para registrar un sistema de IA, recogida en el Anexo VIII del Reglamento. Esta hoja debería incluir también la documentación técnica detallada en el anexo IV.
- Un registro de acceso público que explicite el propósito del sistema de IA, los valores que rigieron su entrenamiento, detalles sobre los conjuntos de datos usados y las estructuras de gobernanza de la organización responsable.

Dos aspectos destacan positivamente de este enfoque. En primer lugar, la exhaustividad de la revisión interna permite una verificación integral de todo el ciclo de vida del sistema de IA,

posibilitando adaptar, a tiempo real, el sistema de gestión de calidad a los estándares requeridos.

En segundo lugar, la publicación de los resultados es una excelente forma de asegurar una rendición de cuentas adecuada a pesar de la falta de personal externo, ya que contribuiría a reducir problemas de información asimétrica entre proveedores y usuarios (Askell et al., 2019). Estos registros podrían ser un buen ejercicio de preparación para la adaptación a los certificados surgidos de la inminente estandarización europea. La industria española ya ha mostrado indicios alentadores de autorregulación, con un certificado de transparencia algorítmica impulsado por Adigital (M. Jiménez, 2022).

Además, recomendamos que estas auditorías no sean vistas como un control interno ordinario, sino como una función empresarial separada con un equipo propio que debe rendir cuentas a la junta directiva, de forma parecida a las auditorías financieras (Schuett, 2023a) (ver entrevista Markus Anderljung - *Apéndice 1*). De igual manera, apoyamos la iniciativa de AESIA de facilitar herramientas digitales que ofrezcan análisis automáticos de código y conjuntos de entrenamiento para realizar autoevaluaciones (Jiménez, 2023). Estos esfuerzos resultan valiosos para reducir costes y universalizar prácticas.

Simulación de ataques (red teaming) y otros escenarios

El *red teaming* es un esfuerzo estructurado para encontrar defectos y vulnerabilidades en un plan, organización o sistema técnico, a menudo desempeñado por un “equipo rojo” que adopta la mentalidad y los métodos de un atacante (Brundage et al., 2020). Esta práctica, muy extendida en el campo de la ciberseguridad, ha empezado a ser utilizada con éxito para anticipar riesgos vinculados a los sistemas de IA, sobre todo en el caso de los modelos de lenguaje (Ganguli et al., 2022). Este fue, de hecho, uno de los ejercicios principales que se realizaron durante el desarrollo de GPT-4 (OpenAI, 2023). En ese caso, la empresa afirma haber identificado riesgos emergentes, lo cual motivó la investigación en seguridad técnica y la implementación de políticas de mitigación que, en muchos casos, redujeron el riesgo. También pudimos observar un cierto consenso entre los expertos entrevistados sobre el uso del *red teaming* como una buena práctica para identificar riesgos sobre la IA (ver entrevistas Toni Lorente, José Hernández-Orallo y Risto Uuk - *Apéndice 1*).

Para el diseño de la simulación de ataques, el elemento principal a tener en cuenta es qué tipo de situaciones se pretende elicitar. Recomendamos que los ejercicios que se lleven a cabo tengan en consideración tres objetivos:

1) Descubrir funcionalidades que propicien usos indebidos. Para los actuales modelos de lenguaje punteros, por ejemplo, resulta particularmente importante identificar comportamientos que puedan ser aprovechados por actores malintencionados para causar daño. Volviendo al caso de GPT-4, estos incluyeron la difusión de discursos de odio, contenido sesgado, desinformación e instrucciones para fabricar armas o perpetrar ciberataques. En este caso, es clave contar con especialistas en un amplio abanico de disciplinas, como la química, la física nuclear, la ciberseguridad, la

economía, el derecho, la sanidad o la educación. Para casos específicos, se podría considerar también la automatización de estas prácticas (E. Perez et al., 2022).

2) Descubrir vulnerabilidades en la infraestructura esencial. El *output* generado por la IA podría ayudar a explotar debilidades en diversas áreas críticas para la seguridad nacional. (Ord et al., 2021) recomiendan la creación de un equipo de expertos que simulen diversos escenarios, como un gran ciberataque a la infraestructura nacional, la liberación de un virus o la interrupción de los servicios de Internet por un periodo extendido de tiempo.

3) Descubrir riesgos estructurales. En este caso, la metodología implicaría simular una serie de escenarios en los que los diversos riesgos estructurales mencionados anteriormente se hayan materializado. Inspirándose en Seger et al. (2020), los responsables de estos ejercicios podrían seguir una estrategia pre-mortem, en la que se presupone que se ha llegado al resultado final y se hace una retrospectiva para descubrir todos los potenciales caminos que podrían llevar a él. El desglose en pasos intermedios permite una mayor concreción a la hora de identificar vulnerabilidades y, por lo tanto, dictaminar qué intervenciones son las más convenientes.

En España, **recomendamos que los organismos públicos se coordinen para institucionalizar estos procesos, creando una red de profesionales independientes enfocados en la identificación de riesgos y ensayo de respuestas.**

Para la exploración de potenciales usos indebidos, recomendamos que INCIBE establezca una red de simulación de ataques efectuados por académicos y especialistas en sectores relevantes, inspirándose en el ejemplo de OpenAI. Estos profesionales deberían ser remunerados, aprobar un test psicotécnico y vincularse a un estricto contrato de confidencialidad. La cooperación público-privada sería especialmente provechosa para distribuir costes y compartir información entre los diversos actores del ecosistema, así como para garantizar que la práctica se estandarice transversalmente con independencia de los intereses y posibilidades de cada actor.

A la hora de descubrir potenciales vulnerabilidades en la infraestructura esencial, el Mando Conjunto del Ciberespacio debería liderar los ejercicios en colaboración con los órganos ministeriales que correspondan. Por ejemplo, se podrían realizar simulación de ataques a centrales eléctricas con la participación de la Subdirección General de Calidad y Seguridad Industrial.

En la Unión Europea, existen ya ejemplos de institucionalización del *red teaming*. En 2018, el Banco Central Europeo adoptó el llamado Red Teaming Ético contra Amenazas (TIBER-EU), un marco para coordinar estos esfuerzos entre países y reforzar la ciber resiliencia del sector financiero europeo. En España, esta iniciativa se ha adaptado a través del hub TIBER-ES, que INCIBE ha decidido aplicar también a otros sectores (INCIBE, 2023).

Sistema de gestión de riesgos

El *Artículo 9 (2), apartado a)*, establece que el sistema de gestión de riesgos debe iniciarse con “la identificación y el análisis de los riesgos conocidos y previsibles vinculados a cada sistema de IA de alto riesgo”. Posteriormente, deberán adoptarse medidas oportunas de gestión de riesgo para eliminar los riesgos en la medida de lo posible o mitigar aquellos que no puedan eliminarse. El sistema de gestión de riesgos es, entonces, un proceso a repetir hasta que todos los riesgos identificados sean aceptables.

La actual presentación de este requerimiento necesita definiciones más claras en dos ámbitos. En cuanto a la primera fase, no se estipula qué se consideran “riesgos conocidos y previsibles”. Schuett (2023a) propone definir “conocido” como aquello que la organización debería conocer tras un esfuerzo razonable, y “previsible” como aquello que no ha ocurrido pero ya puede ser identificado. Aquí, el autor emplea la definición de conocimiento constructivo, que en Derecho se refiere al conocimiento que se presume que uno debería tener al asumir un mínimo deber de diligencia. En este aspecto sirve la teoría de la responsabilidad, que busca la previsibilidad como elemento estructural de diligencia. Se es diligente en la medida en que se realizan acciones para prever los riesgos y evitarlos o gestionarlos correctamente para que no causen daño.

En ese sentido, existen dos problemáticas. Primero, el Reglamento no define qué constituye un nivel razonable de diligencia, por lo que los desarrolladores podrían eludir sus obligaciones alegando desconocimiento. Segundo, debería aclararse hasta qué punto se deben reducir, mitigar o controlar los riesgos. El objetivo de este proceso iterativo es asegurar que todos los riesgos residuales –aquellos que permanecen tras la adopción de medidas– son aceptables (Schuett, 2023a). En este caso, determinar qué riesgos son aceptables implica difíciles juicios normativos y una alta incertidumbre empírica.

Para esto, **recomendamos que el *sandbox* de España se centre especialmente en reforzar la implementación del sistema de gestión de riesgos.** En concreto, sugerimos que los proveedores realicen simulaciones de escenarios adversarios (red teaming) y otros ejercicios de identificación de riesgos, como el análisis modal de fallos y efectos. Este procedimiento se ha utilizado desde hace décadas en la ingeniería de la seguridad para identificar fallos potenciales en un sistema, y ya se ha planteado utilizarlo para la IA (Li & Chignell, 2022).

De igual manera, destacamos la importancia de determinar qué prácticas son necesarias para superar el problema en cuestión. Aquí, los mecanismos clave implican ajustar la arquitectura, los datos, las tareas de entrenamiento o las técnicas de alineamiento para evitar el riesgo (Shevlane et al., 2023). A la espera de estándares más definitorios, la AESIA deberá dictaminar si estos niveles de diligencia y responsabilidad son necesarios para una correcta prevención de los riesgos. La inclusión de buenas prácticas en los informes presentados a las autoridades europeas deberá ser un elemento diferencial para mejorar las provisiones del Reglamento.

Base de datos de incidentes y riesgos

El sistema de gestión de riesgos y el seguimiento posterior a la comercialización previstos en el Reglamento europeo son dos fases fundamentales para evaluar el impacto potencial y tangible de los sistemas de IA. En esta sección, **recomendamos el establecimiento de redes de buenas prácticas para que las lecciones aprendidas en estas etapas sean compartidas con el resto de actores del ecosistema**. Todos los proveedores deberán analizar sistemáticamente estos resultados para alimentar su propio sistema de gestión de riesgos.

Como componente clave de esta red, recomendamos el establecimiento de una **base de datos de incidentes** y una **base de datos de riesgos**, ambas anonimizadas y analizadas por las autoridades nacionales e internacionales con el objetivo de elaborar informes anuales de acceso público que sirvan para el aprendizaje colectivo.

- Base de datos de incidentes

La correcta identificación de incidentes causados por la tecnología ayuda a evitar los mismos fallos o versiones más extremas de estos en iteraciones futuras. Sin embargo, en la mayoría de ocasiones, este aprendizaje se ve restringido a la experiencia individual, ya que los desarrolladores están incentivados a mantener una buena imagen pública y, por lo tanto, a ocultar los percances en los que están implicados (Brundage et al., 2020). Probablemente, la solución a este problema pasa por la creación de canales cooperativos para compartir esta información sin comprometer la reputación de los afectados.

El *Artículo 62* ya ordena a los proveedores que notifiquen cualquier incidencia grave o fallo de funcionamiento de sus respectivos sistemas de IA. Este aviso deberá presentarse a la correspondiente autoridad de vigilancia del mercado, que posteriormente informará al organismo supervisor del Reglamento. Por otro lado, el *Artículo 60* estipula que la Comisión deberá crear y mantener una base de datos de acceso público que contenga información sobre los sistemas de IA de alto riesgo registrados en la UE. La información solicitada, recogida en el Anexo VIII, incluye los datos del proveedor, una descripción de la finalidad del sistema, una lista de países donde el sistema se haya puesto en servicio y una copia de la declaración UE de conformidad, entre otros.

En este contexto, **recomendamos que todos los incidentes graves[5] notificados de acuerdo con el *Artículo 62* sean recopilados sistemáticamente en una base de datos paralela a la prevista en el *Artículo 60***. Para evitar conflictos, las autoridades nacionales y europeas deberían anonimizar los incidentes, garantizando que la relación entre incidente y responsable se mantenga confidencial. Una vez establecida la base de datos, un equipo de especialistas debería dedicarse a su análisis para encontrar pautas comunes y extraer lecciones. Esta recomendación ya fue impulsada por varios grupos en su respectivo

[5] De acuerdo con la definición propuesta en el Reglamento, entendemos que son “incidentes graves” aquellos que, directa o indirectamente, causen daños graves para la integridad física, propiedades o entorno de una persona, así como alteraciones graves e irreversibles de la gestión y el funcionamiento de infraestructura crítica.

feedback a la propuesta de la Comisión (Future Of Life Institute, 2021) (Clarke et al., 2021). También ha sido apoyada por expertos durante nuestro proceso de consultas (ver entrevista Risto Uuk y Toni Lorente, *Apéndice 1*)

Un buen ejemplo para esta base de datos podría ser el AI Incident Database (AIID) de Partnership on AI, una compilación de daños o cuasi daños causados por el despliegue de sistemas de IA. Este recurso, inspirado en otros sectores como la aviación o la seguridad informática, tiene como objetivo facilitar un aprendizaje basado en la experiencia para prevenir y mitigar futuros incidentes.

El *sandbox* en España presenta una buena oportunidad para que los organismos públicos prueben la recopilación sistemática de incidentes causados por sistemas de IA. Como principal receptor de las notificaciones, AESIA debería ser responsable de su recopilación. Asimismo, el ejercicio debería contar con la participación de instituciones dedicadas a la ciberseguridad. El Equipo de Respuesta ante Emergencias Informáticas (CERT) de INCIBE gestiona un repositorio de acceso público con más de 75.000 vulnerabilidades de seguridad en sistemas tecnológicos. Estos registros se basan principalmente en la lista internacional de CVE (Common Vulnerabilities and Exposures), que facilita el intercambio de información –problemas y soluciones– entre organizaciones y países.

Para una mayor comprensión y coherencia, es importante que estas bases de datos nacionales estén relacionadas entre sí. Esta interconexión permitirá a los investigadores y desarrolladores acceder a una mayor cantidad de información sobre incidentes, lo que fortalecerá el proceso de aprendizaje de la IA y reducirá el riesgo de incidentes futuros (ver entrevista Toni Lorente - *Apéndice 1*).

A nivel comunitario, el Comité Europeo de Inteligencia Artificial podría facilitar esta coordinación y agrupación. Esta recomendación está en consonancia con el *Artículo 58*, que prevé que una de las funciones del Comité sea recopilar y compartir conocimientos técnicos y buenas prácticas entre los Estados miembros. En un esfuerzo para ampliar las competencias del Comité, el Consejo propuso en su Postura Común que el *Artículo 58* incluyera también la promoción y apoyo de investigaciones transfronterizas de vigilancia del mercado (Consejo, 2022). El resultado final podría asemejarse al trabajo realizado por ENISA, la agencia europea de ciberseguridad, a través del Sistema de Notificación y Análisis de Ciberseguridad (CIRAS). Este organismo recopila, anonimiza y analiza los datos enviados por las autoridades nacionales para elaborar un informe anual que recoge las principales lecciones.

- Base de datos de riesgos

Teniendo en cuenta el impacto potencial de la IA, un enfoque reactivo podría ser insuficiente a largo plazo. Vincular el aprendizaje a la respuesta ex-post a incidentes puede establecer un precedente peligroso, ya que la gravedad de los incidentes probablemente escalará con las capacidades de la IA. La mera posibilidad de que un incidente cause un daño irreparable es motivo suficiente para que los actores que contemplen el riesgo compartan sus observaciones antes de que dicho incidente se materialice.

Los responsables del AIID proponen explorar una clasificación en dos categorías: incidentes y problemas (McGregor et al., 2022). Esta última se referiría a “daños causados por un sistema de IA que aún tienen que ocurrir o ser detectados”. Esta taxonomía iría en concordancia con el CVE, que también distingue entre “eventos” y “riesgos”.

En este sentido, **recomendamos que las bases de datos incluyan también los riesgos conocidos y previsibles** que, en caso de materializarse, puedan provocar un incidente grave. Los ejercicios de evaluación de modelos y simulación de ataques propuestos en las secciones anteriores deberían alimentar estos esfuerzos. Así pues, la identificación y recopilación de riesgos debería ir a cargo de los organismos evaluadores, quienes transmitirían los principales hallazgos a las autoridades nacionales.

En el mismo sentido que las bases de datos de incidentes, es importante considerar la conexión entre los listados de riesgos y la interacción con otras bases de datos mundiales (ver entrevista Toni Lorente - *Apéndice 1*).

Responsabilidad del proveedor en la integridad del sistema de IA

El *Artículo 28(1)* estipula que cualquier distribuidor, importador, usuario o tercero será considerado proveedor si comercializa el sistema de IA con su nombre comercial o si modifica sustancialmente características como su finalidad prevista. Esta provisión resulta fundamental para exigir responsabilidades legales a aquellos actores malintencionados que empleen sistemas de IA en contradicción con sus instrucciones de uso.

Por otra parte, el *Artículo 28(2)* establece que, cuando el sistema de IA haya sido modificado sustancialmente, incluyendo cambios en su finalidad prevista, el proveedor que inicialmente lo introdujo en el mercado dejará de ser considerado proveedor a efectos del Reglamento. Este párrafo ha sido discutido porque podría suponer laxitud en las obligaciones que pesan sobre el despliegue del sistema de IA por parte del proveedor original.

El Consejo, a través del *Artículo 23a*, enmendó el *Artículo 28* de la propuesta de la Comisión. El cambio más significativo es que la modificación de la finalidad prevista del sistema, cuando esta suponga que el sistema pasa a ser de alto riesgo, se elimina de la lista de escenarios en los que el proveedor original deja de ser considerado legalmente el proveedor. El Parlamento propuso eximir al proveedor original de las obligaciones vinculadas a modificaciones sustanciales, si bien se le requiere que facilite al nuevo proveedor toda la información y documentación del sistema para facilitar el cumplimiento con el Reglamento.

Recomendamos que la responsabilidad legal de los proveedores originales se mantenga siempre que se produzca una modificación de la finalidad prevista o que cualquier uso del sistema de IA suponga un riesgo inaceptable, independientemente de si este uso entra o no en contradicción con las instrucciones del sistema. Esto incentivaría a los proveedores a asegurar sus sistemas para evitar modificaciones, y a garantizar que las instrucciones de uso no puedan ser esquivadas para causar daño. Recomendamos que la responsabilidad legal de los proveedores se extienda para incluir

también los daños causados por reproducciones e imitaciones de un modelo original que se haya filtrado o extraído forzosamente gracias a la falta de medidas de seguridad.

Idealmente, el Reglamento debería ir acompañado de un trabajo adicional para universalizar contratos que ayuden a reforzar el control sobre toda la cadena de valor de la IA. Inspirándose en el caso de OpenAI, estos controles podrían incluir, entre otros, un proceso de revisión para aprobar el uso de la API, límites al número de interacciones con la API y el monitoreo de datos para verificar posibles usos indebidos. Asimismo, los proveedores deberían desplegar el sistema de forma gradual y asegurando la confidencialidad del proceso. El *sandbox,* que dará un peso importante a las pymes, será un entorno ideal para probar la coordinación de estas con los proveedores originales.

Planes para la intervención en productos dañinos

El *Artículo 65* del Reglamento ordena a las autoridades de vigilancia del mercado que, cuando un sistema de IA presente un riesgo inaceptable, adopten las medidas correctoras oportunas para adaptar el sistema de IA a los requisitos del Reglamento o bien retirarlo del mercado en un plazo proporcional a la naturaleza del riesgo. Cuando el operador sea incapaz de ello, será la autoridad de vigilancia la que adopte estas medidas. Esta provisión está en línea con el *Artículo 20* del Reglamento (CE) Nº 765/2008 y el *Artículo 19* del Reglamento (UE) 2019/1020, que estipulan que las autoridades de vigilancia del mercado garantizarán que los productos que planteen un riesgo grave sean recuperados o retirados, o que se prohíba su comercialización.

Sin embargo, esta regulación no detalla cómo se pueden llevar a cabo estas acciones y, sobre todo, cómo se debe desplegar un sistema de modo que sea factible su retirada en caso de ser necesario (véase la entrevista a Charlotte Siegmann - *Apéndice 1*).

Detener la comercialización es un desafío presente en varios sectores. El Observatorio de Vigilancia de Mercado (UNE, 2022) apunta que la retirada de productos inseguros se ve dificultado por la amplitud de las fronteras comunitarias, la falta de recursos suficientes por parte de las autoridades de vigilancia, los amplios plazos de tramitación, la falta de actuaciones automáticas y la irrupción del comercio en línea sin definición legal de las responsabilidades. Por su parte, la SETELECO destaca en el *Plan sectorial para la vigilancia del mercado de equipos de Telecomunicación* que muchos productos no conformes provienen de terceros países por comercio electrónico, y que la falta de trazabilidad en la cadena de suministro crea dificultades a la hora de requerir medidas correctoras (SETELECO, 2022). Estas dificultades podrían verse exacerbadas en el caso de la IA, puesto que su difusión resulta especialmente difícil de controlar e involucra potencialmente a numerosos terceros.

Recomendamos que la SEDIA elabore un Plan sectorial para la vigilancia del mercado de sistemas de Inteligencia Artificial, inspirándose en el documento análogo para los equipos de telecomunicación. Este proyecto debería consistir en una planificación de campañas proactivas en tres fases:

1. Planificación de la campaña, incluyendo estudios de mercado y evaluación de los riesgos asociados a cada sistema de IA. Recomendamos que los sistemas de alto riesgo sean sometidos a campañas específicas y que los modelos fundacionales deban superar campañas de control sistemáticas.
2. Ejecución de la campaña:
    a. Inspecciones visuales: comprobaciones del funcionamiento del sistema en la nube o en el hardware del proveedor, según proceda.
    b. Inspecciones documentales: evaluación de la documentación técnica requerida en el *Artículo 11* del Reglamento europeo para la IA.
    c. Inspecciones con retirada: cese temporal de la comercialización del sistema de IA, puesto a disposición de los laboratorios de ensayos para la comprobación de los requisitos administrativos y técnicos. Aplicable solamente a aquellos sistemas que impliquen riesgos salientes para la salud, la seguridad o la protección de los derechos fundamentales de las personas.
3. Análisis de resultados y ejecución de medidas. De acuerdo con el Artículo 65 del Reglamento, las autoridades de vigilancia del mercado podrán adoptar medidas correctoras y, cuando no sea posible, prohibir o restringir la comercialización.

En cuanto a las obligaciones impuestas al proveedor, el control sobre la cadena de valor estipulado en la recomendación anterior vuelve a ser importante para garantizar una actuación rápida en caso de necesidad. En este sentido, los proveedores deberán analizar concienzudamente los archivos de registro generados por el funcionamiento del sistema de IA, de modo que se puedan detectar incidencias graves a tiempo real. En la previsión de estos casos, los proveedores deberán reservar en los términos de uso de sus APIs el derecho a cortar el servicio.

Sistemas de propósito general y modelos fundacionales

El principal punto de partida de la propuesta de ley de la Comisión es la clasificación de los sistemas de IA según el nivel de riesgo que implican. Actualmente, esta clasificación vincula el riesgo a los juicios éticos de la tecnología y a sus ámbitos de aplicación. Sin embargo, como se ha explicado, el desarrollo de sistemas avanzados conlleva una serie de riesgos intrínsecos que los reguladores y decisores políticos deben considerar. En este contexto, es importante considerar que los sistemas de propósito específico tienen un mercado y riesgos más acotados, mientras que los sistemas de propósito general están sujetos a mayor variación (ver entrevista José Hernández-Orallo - *Apéndice 1*).

Esta cuestión está siendo objeto de intensos debates en el seno de la Unión Europea. El Consejo propuso aplicar los requerimientos de los sistemas de alto riesgo a los sistemas de IA de propósito general (IAPG) (Consejo, 2022). Por su parte, el borrador del Parlamento distingue entre un sistema de IAPG y un modelo fundacional. El primero se define como “un sistema de IA que se puede usar y adaptar a una amplia gama de aplicaciones para las que no fue diseñada intencional y específicamente”, mientras que el segundo se entiende como “un modelo de IA que se entrena en amplios conjuntos de datos a escala, diseñado para la generalidad de la producción y adaptable a una amplia gama de tareas”. Con base en estas

definiciones, esperamos que todos los sistemas punteros definidos por medidas computacionales coincidan en gran medida con los modelos fundacionales.

Recomendamos que tanto los sistemas de IAPG como los modelos fundacionales sean incluidos explícitamente en el texto legislativo. En el caso de los modelos fundacionales, los desarrolladores deberían asumir tanto las obligaciones requeridas para los sistemas de alto riesgo como las definidas en este informe, con especial énfasis en las evaluaciones del modelo. Para los sistemas de propósito general, recomendamos que se apliquen las primeras y se consideren las segundas, especialmente los ejercicios de *red teaming*.

Además, existen varias consideraciones que son importantes para los sistemas de propósito general y que no se observan en las regulaciones de sistemas más específicos (ver entrevista Markus Anderljung - *Apéndice 1*):

- La relación y distribución de responsabilidades entre aquellos que desarrollan modelos de IAPG y aquellos que adaptan estos modelos para usos específicos, como se ha explicado anteriormente.
- La colaboración entre los desarrolladores de modelos de IAPG y las autoridades competentes u otros actores externos para identificar y prevenir riesgos y usos indebidos de la tecnología. Recomendamos que el sistema de gestión de riesgos sea especialmente estricto para sistemas de IAPG, de modo que se prevean todos los posibles usos indebidos. Esto se podría hacer a través de ejercicios de simulación de ataques*,* que debería ser normativo para casos especialmente delicados como el reconocimiento facial.

Por último, es importante tener en cuenta que el éxito del *sandbox* vendrá determinado por el alcance que se le dé y el marco regulatorio que se pruebe. Si no se ensaya el acto de implementación de los sistemas de propósito general, las conclusiones que se den en el *sandbox* no necesariamente serán extensibles a dicha tecnología (ver entrevista Toni Lorente - *Apéndice 1*).

Sector militar

De acuerdo con el *Artículo 2* del Reglamento, los sistemas de IA desarrollados o utilizados exclusivamente con fines militares están explícitamente excluidos del ámbito de aplicación. Esto se justifica manifestando que, cuando su uso sea competencia exclusiva de la política exterior y de seguridad común regulada en el título V del Tratado de la Unión Europea (TUE), no estará cobijado por el Reglamento.

Si bien comprendemos la falta de competencias de la Unión Europea en el sector, **recomendamos elaborar pautas y directrices para guiar la aplicación del derecho internacional humanitario en los usos militares de la IA.**

En este sentido, destacan dos resoluciones del Parlamento Europeo. En 2018, los eurodiputados abogaron por iniciar negociaciones internacionales para desarrollar un instrumento vinculante que prohíba las armas autónomas letales (2018/2752(RSP)). Y en 2021, el Comité de Asuntos Legales publicó un conjunto de pautas para la interpretación del derecho internacional, entre las que destaca la necesidad de garantizar la supervisión y responsabilidad humanas en el uso de la IA (2021/C 456/04).

Es importante que España y la Unión Europea vuelvan a erigirse como líderes para el mantenimiento de la paz a nivel global. Esto incluye seguir empujando para la prohibición de las armas letales autónomas, así como la automatización del comando y control nuclear. Pedimos asimismo que la excepción militar en el Reglamento no genere salvedades innecesarias, y que los proveedores se aseguren de que sus sistemas de IA no se integren en aplicaciones militares inaceptables.

Conclusión

El campo de la inteligencia artificial ha experimentado un rápido avance en los últimos años, impulsado principalmente por los desarrollos en el aprendizaje automático y el aumento en el poder computacional. Existen grandes expectativas en cuanto a la posibilidad de desarrollar IA que posea habilidades cognitivas de dominio general, lo que podría tener un impacto significativo en una amplia gama de áreas de aplicación; sin embargo, también surgen preocupaciones sobre los riesgos asociados a su desarrollo e implementación. En este informe, se han identificado dos categorías principales de riesgos asociados a la IA: riesgos adversarios, que pueden derivarse de un uso indebido de los sistemas de IA o del desarrollo de sistemas avanzados desalineados; y riesgos estructurales, asociados al despliegue a gran escala de la tecnología.

En respuesta a estos desafíos, se han propuesto diversas medidas regulatorias, como el Reglamento Europeo para la IA, que busca establecer normas armonizadas y requisitos para el uso de sistemas de IA en sectores críticos. En este contexto, España tiene una oportunidad única para contribuir positivamente al desarrollo de la regulación de la IA, a través de su participación en un *sandbox* regulatorio que permitirá probar la viabilidad del Reglamento y explorar políticas adicionales que refuercen sus objetivos, al tiempo que consolidan la conciencia pública sobre los riesgos y beneficios de la IA.

En el contexto de la inminente regulación europea, y aprovechando la situación privilegiada que tendrá España para influenciar su implementación, hemos presentado siete políticas que consideramos que ayudarán a mejorar la gobernanza de la IA: el seguimiento de concentraciones del cómputo, la realización de auditorías, la ejecución de ejercicios de *red teaming*, el refuerzo de los sistemas de gestión de riesgos, la elaboración de una base de datos de incidentes, la implementación de mecanismos de control a lo largo de la cadena de valor y el desarrollo de planes de intervención ante emergencias. De igual forma, hemos expuesto dos sugerencias para asegurar que el Reglamento y otros procesos legislativos futuros abarcan realmente los sistemas de IA que comportan un mayor riesgo. Todas estas

recomendaciones conforman un marco que España puede adoptar para establecer buenas prácticas en su esfuerzo pionero para gobernar y regular la IA.

Autoría del informe

Nombre		Afiliación
Guillem	Bas Graells	Riesgos Catastróficos Globales
Roberto	Tinoco	Riesgos Catastróficos Globales
Jaime	Sevilla Molina	Riesgos Catastróficos Globales, Epoch, Centre for the Study of Existential Risk (Cambridge University)
Jorge	Torres Celis	Riesgos Catastróficos Globales
Mónica	Ulloa Ruiz	Riesgos Catastróficos Globales
Daniela	Tiznado	Riesgos Catastróficos Globales

Agradecimientos

Agradecimientos especiales por su ayuda y comentarios a Pablo Villalobos, Pablo Moreno, Toni Lorente, José Hernández-Orallo, Joe O'Brien, Rose Hadshar, Risto Uuk, Samuel Hilton, Charlotte Siegmann, Javier Prieto, Jacob Arbeid y Malou Estier.

Apéndices

Apéndice 1. Resúmenes de entrevistas

Toni Lorente
Associate, AI Governance, The Future Society (TFS)

Toni Lorente plantea que la regulación europea debe abordar los sistemas de propósito general. En la actualidad, lo que propone el texto es la regulación de los riesgos asociados a ciertos usos en distintos ámbitos. Esto no es necesariamente negativo, dado que es agnóstico en cuanto a la tecnología.

El éxito del *sandbox* y, sobre todo, el éxito en relación a ciertas tecnologías -como los sistemas de propósito general- vendrá determinado por el alcance que se le dé y el marco regulatorio que se pruebe. Si no se ensaya el acto de implementación de los sistemas de propósito general, las conclusiones que se den en el *sandbox* no necesariamente serán extensibles a dicha tecnología.

En un mismo sentido, para el desarrollo del *sandbox* de España es esencial que se realice un proceso de divulgación tanto sobre su finalidad como su alcance a los diferentes actores; por otro lado, la legitimidad del *sandbox* vendrá dada por la inclusión de distintos elementos de gobernanza, no solo la ley, y de todos los actores interesados. Por ejemplo, entender las interacciones entre normas, estándares, regulación ya existente en materia de protección de datos y el nuevo marco legal facilita una futura armonización; es también importante que exista representación de todos los actores en la gobernanza. Algunos sectores, como los laboratorios de sistemas de propósito general podrían involucrarse más.

Otros aspectos importantes tienen relación con mecanismos de cumplimiento, bases de datos de incidentes y gestión de la información. Sobre los mecanismos, existen varios para que se cumpla con la regulación sin poner en riesgo assets de las compañías, como la propiedad intelectual sobre su tecnología; también es importante reforzar la elaboración de bases de datos de incidentes. Ya existe una base de la OCDE, pero hace falta analizar cómo extraer aprendizajes de cada incidente, especialmente sobre gobernanza de la IA; y por último, otro aspecto importante es la gestión de las asimetrías en la información, especialmente en las etapas de investigación y desarrollo, así como el desarrollo y gestión de certificaciones y estándares.

Es relevante considerar el impacto del efecto Bruselas a escala global, tanto en relación a la interoperabilidad de estándares y normas, como en la dinamización del mercado en un contexto global.

Pablo Villalobos
Staff Researcher, EPOCH

Existen drivers primarios y secundarios en el desarrollo de la IA. Los primeros son aquellos factores que influyen en el modelo directamente, como algoritmos, datos y cómputo. Los segundos son indirectos al modelo, como el capital humano y el financiamiento.

En cuanto a las mejoras algorítmicas, se puede observar a nivel cualitativo cuáles técnicas nuevas se han implementado. Es razonable esperar que conforme los algoritmos se vuelvan más generales, haya menos desarrollo de algoritmos completamente nuevos. Habrá mejoras considerables pero no con cambios muy profundos.

Si sigue aumentando el uso de datos al mismo ritmo, parece bastante probable que se usen todos los datos disponibles, ya que estos crecen más lento. Hay varias técnicas que pueden usarse para tomar menos datos o usar datos “sintéticos”. También puede llegar antes otro límite relacionado con la financiación, pues no todo el mundo puede mantener el ritmo de inversión que se está teniendo en la actualidad.

En cuanto a capacidades futuras, Villalobos plantea que con los sistemas actuales se puede llegar a la automatización de tareas concretas. Si se tiene en cuenta la mejora en los modelos en la próxima década, se mejorará en todo lo que pueda ser probado varias veces para que los fallos no sean críticos, sobre todo trabajos digitales. Lo que no se ve probable es que se realicen coches autónomos y actividades relacionadas con la construcción.

En cuanto a gestión de datos, sugiere trabajar sobre la eficacia del entrenamiento y no tanto en reducir la desinformación (esto último es más complejo), robustecer el origen de los datos (sólo usar papers científicos, artículos oficiales), usar modelos ya entrenados para filtrar, eliminar duplicados, quitar contenido dañino, y por último, realizar feedback automatizado y probarlo en una cantidad de situaciones y tener seres humanos etiquetando los resultados.

Para equilibrar los progresos con los riesgos potenciales de la IA, indica que se pueden utilizar mecanismos ya existentes como los impuestos progresivos con beneficios sociales. Se ha visto cómo la población activa disminuye por las cortas ventanas de adaptación y esto puede llevar a la implementación de la renta básica.

España tiene bastantes empresas de IA pero no enfocadas a progresos más generales. El país se verá beneficiado al implementar dichos progresos para su economía, pero esto depende de cómo se concentran las ganancias de estos sistemas, pues puede que se vayan a empresas extranjeras y no dejen beneficios considerables.

Risto Uuk

Policy Researcher, Future of Life Institute (FLI)

La ley de IA de la UE tiene varios objetivos clave, como ayudar al mercado de la UE a funcionar mejor, evitar fricciones entre los estados, promover la IA y hacer que la ley sea aplicable a todos los países miembros. Las discusiones pueden ser largas y es necesario encontrar un equilibrio y un compromiso. Algunos responsables políticos esperan finalizar las negociaciones para finales del año, mientras que otros esperan que se retrase hasta el nuevo año.

El avance de la IA en Europa también se beneficiaría de una visión clara de los incidentes de seguridad a nivel europeo, ya que esto facilitaría el análisis de qué investigación o

regulación puede ser necesaria a medida que surgen tendencias en el mercado único. Por eso, recomendamos que los Estados miembros también informen sobre los incidentes de seguridad a una base de datos de la UE. La UE debería considerar abrir el acceso a los "sandboxes" a las PYME de fuera de la Unión. Esto promovería la difusión de los estándares de la UE a nivel mundial.

Algunas de las discusiones sobre la ley de IA tendrán lugar en el Consejo de IA de la UE, que ayuda a la Comisión Europea a evaluar cómo está funcionando la regulación. También se está discutiendo la creación de una agencia de IA más grande y poderosa que tenga su propia entidad jurídica.

En cuanto a la cuestión de las auditorías, se señala que se espera que las empresas realicen principalmente sus propias evaluaciones, pero esto creará problemas de confianza porque no se puede confiar plenamente en la autoinformación. El desarrollo adicional de las auditorías de terceros puede ser bienvenido. Los ejercicios de *red teaming* son una buena idea para encontrar vulnerabilidades, ya sea que la auditoría sea interna o externa.

También existe la posibilidad de "safety-washing", lo que significa que las empresas afirman que están trabajando en la seguridad de la IA sin tomar medidas significativas en la práctica. Es importante informar no solo sobre los incidentes, sino también sobre los incidentes cercanos, que son oportunidades de aprendizaje. Los ejercicios de *red teaming* no deben ser públicos, pero los incidentes reales sí. La ciberseguridad puede ofrecer aprendizajes útiles sobre riesgos y buenas prácticas.

Se necesita una metodología sólida para evaluar los riesgos de la IA, ya que actualmente se hace principalmente en función de juicios intuitivos. Los problemas sociales a gran escala relacionados con, por ejemplo, la IA y la democracia, el Estado de derecho y el medio ambiente deben ser considerados, pero llevan más tiempo evaluarlos y aún no están bien definidos.

José Hernández-Orallo

Catedrático en la Universitat Politècnica de València (UPV)

Hasta hace poco, se consideraba que el aprendizaje profundo (Deep Learning) consistía en una combinación de algoritmos, datos y poder de cómputo. Sin embargo, esta perspectiva se ha vuelto obsoleta debido a una convergencia en el campo. En la actualidad, se reconoce ampliamente que sólo unos pocos algoritmos son capaces de resolver una amplia gama de tareas. Este enfoque ha ampliado el acceso a la inteligencia artificial, permitiendo que más personas utilicen sistemas de propósito general como GPT.

A pesar de los esfuerzos por abordar el sesgo en los sistemas de inteligencia artificial mediante filtros y controles, persisten problemas de sesgo latentes. Por ejemplo, una palabra puede tener connotaciones distintas que afectan las respuestas del sistema. Esto plantea el dilema entre tener sistemas menos generales y más previsibles, o más generales y menos previsibles. La generalidad de estos sistemas implica cierta imprevisibilidad, similar a la imprevisibilidad inherente en las interacciones humanas. La clave radica en regular y establecer normas más estrictas para la inteligencia artificial, aunque también preocupa el

posible uso malicioso de esta tecnología. En última instancia, no solo se trata del problema de las máquinas, sino del uso que se les dé.

El sistema GPT tiene un mecanismo de feedback en el que se utiliza el aprendizaje por refuerzo para ajustar las probabilidades de salida del modelo. Este proceso implica la intervención humana para modificar las salidas problemáticas y seleccionar alternativas más adecuadas. Sin embargo, el sistema no tiene un feedback continuo y los pesos originales de la red no se modifican. Aunque se utilizan términos como "filtro" o "aprendizaje por refuerzo", en realidad no se están cambiando los pesos de la red neuronal. El feedback del usuario no es automático y se recoge de forma periódica para reentrenar el sistema y lanzar nuevas versiones.

En el campo de la inteligencia artificial, existe una falta de regulación similar a la que se consolidó en la informática en los años 60 y 70. Esta ausencia de regulación ha llevado a una mentalidad en la que se asume que simplemente se necesita agregar un descargo de responsabilidad al usuario y luego se le permite hacer más cosas. Esta cultura informática ha permeado en muchos centros dedicados a la inteligencia artificial, lo que significa que los usuarios asumen la responsabilidad total de cualquier problema que puedan encontrar. A diferencia de otras áreas, donde existen regulaciones mínimas para proteger a los consumidores, en el ámbito del software, se espera que los usuarios acepten los posibles riesgos y consecuencias sin una regulación clara. Esta falta de regulación ha causado daños significativos y ha generado la necesidad de establecer estándares y normativas adecuadas para el uso de la inteligencia artificial.

A medida que los sistemas de inteligencia artificial adquieren acceso a un vasto conocimiento humano y se escalan rápidamente, es probable que se busquen soluciones para superar los límites de los datos disponibles. Se explorarán enfoques como la generación de datos de entrenamiento, la generación de ejercicios matemáticos resueltos y la generación de datos basados en observación de experimentos científicos. Sin embargo, a pesar de que estos sistemas pueden adquirir un conocimiento casi ilimitado del mundo real, es poco probable que a corto plazo sean capaces de descubrir nuevas leyes físicas, por ejemplo. Se espera que surjan nuevos paradigmas que aprovechen la información infinita disponible, pero es importante reconocer que estos datos no sustituyen el conocimiento acumulado durante milenios a través del lenguaje y la experiencia humana. Aunque los avances en inteligencia artificial continúan, existe la necesidad de nuevas ideas y enfoques para superar los límites actuales y seguir avanzando.

Charlotte Siegmann

Pre-Doctoral Research Fellow en Economía, Global Priorities Institute, Universidad de Oxford

Siegmann plantea que es mejor proponer recomendaciones que tienen una pequeña posibilidad de ser implementadas, pero que de suceder serían muy beneficiosas. Sugiere que la implementación de grandes modelos de lenguaje en burocracias puede ser muy útil en muchos sectores, pero también podría tener consecuencias económicas y de acceso a recursos. Propone que se establezcan regulaciones para las empresas que usan estos modelos de lenguaje, obligándolas a tener planes para deshacerse de ellos en caso de ser

necesario, y estableciendo una agencia reguladora fuerte como la que existe para productos farmacéuticos.

Menciona la importancia de la interpretación de los modelos de inteligencia artificial y la necesidad de realizar pruebas de auditoría para evaluar la seguridad del sistema. Señala que los reguladores europeos podrían no tener la experiencia necesaria para realizar estas pruebas, para lo cual plantea que los auditores deben tener las capacidades necesarias para realizar pruebas efectivas y reducir el riesgo de fugas de información en el proceso de auditoría. Sugiere que la calidad de la burocracia que rodea al modelo podría ser un factor importante en la obtención de resultados precisos e indica que la comunidad de código abierto podría ser un recurso valioso para los auditores en la adquisición de conocimientos especializados.

Samuel Hilton

Research Affiliate, The Centre for the Study of Existential Risk (CSER), Universidad de Cambridge

Hilton señala que las deficiencias en la comprensión de los políticos sobre ciertos temas pueden variar de país a país, particularmente en cuanto a riesgos a nivel nacional. Aunque algunos de los países tienen un sistema de gestión de riesgos, la forma en que se identifica y se presenta el registro de riesgos nacionales a menudo no es efectiva. Tras identificar los riesgos, es crucial actuar para manejarlos, lo que implica responsabilidad y es algo que no se encuentra siempre.

En cuanto a la comunicación de temas hipotéticos, como los riesgos de la IA, es importante ser concreto y evitar caer en escenarios de ciencia ficción. Es recomendable tener una perspectiva a largo plazo sobre los riesgos y ser específico al hablar de ellos. Es crucial captar el interés del político, evitando sonar demasiado fantasioso.

La mayoría de los políticos se enfocan en los problemas cotidianos y lo que aparece en las noticias, por lo que las ideas a largo plazo no suelen ser su prioridad. Para comunicarse con ellos de manera efectiva, es útil centrarse en sus necesidades específicas.

En cuanto a la simulación de adversarios (red teaming) a nivel nacional, se destaca la importancia de la responsabilidad y la necesidad de un sistema de auditoría, así como de ejercicios de escenario que permitan la planificación y capacitación. Es fundamental contar con un plan gubernamental para la gestión de riesgos.

Markus Anderljung

Head of Policy - Research Fellow, GovAI

Anderljung señala que ciertos tipos de requisitos del reglamento pueden ser cumplidos selectivamente en una sola jurisdicción, lo que impide un efecto de facto. La incertidumbre puede surgir cuando no se han establecido medidas de cumplimiento, lo que puede resultar en la selección de algunos productos y servicios, y otros no. Esto es muy importante de considerar y España tendrá un papel importante en esto para definir este alcance.

También es importante señalar que se busca mejorar la regulación y la auditoría. Para lo primero, es importante para tener éxito considerar los sistemas de propósito general. Para lo segundo, la auditoría debe establecer el alcance de las evaluaciones de conformidad.

Al menos para los modelos fundamentales, se deberían exigir auditorías internas y externas obligatorias. Una auditoría interna no es simplemente una auditoría que se realiza internamente en una empresa. La auditoría interna es una función empresarial separada y tiene su propio equipo, quien además le debe rendir cuentas a la Junta directiva y no al CEO (parecido a las auditorías financieras).

En cuanto a las auditorías externas, la evaluación de si un modelo podría ser peligroso requiere de expertos (la experiencia técnica adecuada y correcta) y esto es problemático si estas evaluaciones sólo las realiza un pequeño número de actores. Uno de los temas que el EU AI Act puede mejorar es que debe hacerse explícito que deben existir múltiples actores intentando encontrar fallas en un sistema IA y de manera simultánea e independiente cada uno, lo que proporcionaría una revisión más robusta y permitiría identificar y corregir cualquier problema o deficiencia desde diferentes ángulos.

Es importante identificar los riesgos y tener claro qué se necesita para gestionarlos. Se sugiere que lo más factible es informar a la autoridad antes del despliegue, idealmente con un plazo de al menos tres meses, para que las autoridades correspondientes estén informadas y puedan tomar las medidas adecuadas si es necesario.

Por último, señala tres niveles que para él son importantes, en particular, para los sistemas de propósito general:

1. La relación y distribución de responsabilidades entre aquellos que desarrollan modelos GPAI y aquellos que adaptan estos modelos para usos específicos. Es importante establecer una evaluación de conformidad para los modelos GPAI y gestionar adecuadamente la transferencia de responsabilidades entre las partes involucradas.
2. La colaboración entre los desarrolladores de modelos GPAI y las autoridades competentes en la identificación y prevención de usos de alto riesgo o indebidos de la tecnología. Los desarrolladores de modelos GPAI deben tener responsabilidades adicionales en esta área, ya que pueden estar en una posición particularmente importante para detectar y prevenir riesgos.
3. La gestión de riesgos y la implementación de mecanismos de control y evaluación por parte de actores externos. Este aspecto se considera el más importante en términos de prevención de resultados catastróficos. Implica identificar las características positivas y negativas que un sistema puede tener, evaluar los riesgos asociados con la implementación del sistema y permitir que actores externos revisen estas evaluaciones de riesgos. Estas evaluaciones deben informar cómo se implementa el sistema.

Ricardo Baeza-Yates
Director de Investigación en el Instituto de IA Experiencial de Northeastern University - ex miembro del Consejo Asesor de Inteligencia Artificial de España

En proyectos de gran envergadura como la inteligencia artificial, y que implica a las personas, se destaca la necesidad de un análisis de impacto, potencialmente en términos de derechos humanos, y la necesidad de demostrar que los beneficios superan los daños. Una regulación que requiera un mínimo de competencia técnica, ética y administrativa podría ayudar a prevenir posibles daños sobre las personas.

Se resalta la importancia de la colaboración entre el sector privado y el gobierno para generar incentivos políticos y económicos correctos e impulsar la ética dentro de la inteligencia artificial. Se sugiere que las políticas que permiten al sector privado presentar propuestas directamente al gobierno, como las de Nueva Zelanda y el Reino Unido, podrían aplicarse en España. Señala la necesidad de una mayor colaboración y diálogo entre los sectores público y privado, ya que actualmente hay poca comunicación formal entre ellos.

Se sugiere que, en lugar de certificar modelos, se podría certificar el proceso de su creación, al estilo de las normas ISO 9000. Esta certificación verificaría que el proceso cumple con ciertos estándares, incluyendo la consulta con usuarios antes de la implementación y la realización de pruebas exhaustivas. Los comités éticos también podrían ser parte del proceso para evaluar posibles sesgos o discriminaciones en los modelos -tanto internos como externos-.

En cuanto a la medición de poder computacional para evaluar las capacidades de un modelo y enfocar las auditorías sobre estos, se señala que el impacto ético de la tecnología puede ser significativo, independientemente del uso de recursos computacionales. No obstante, se puede ver desde la utilización eficiente de los recursos energéticos, tanto en entrenamiento como en uso, especialmente con modelos grandes que consumen enormes cantidades de energía durante su uso continuo por millones de personas.

Existe escepticismo respecto a la realización de un piloto en España para una regulación de la Unión Europea que aún no existe, ya que podría resultar en un desperdicio de recursos. Hacer un ensayo basado en reglas que pueden cambiar podría ser una pérdida de tiempo y dinero.

Ibán García del Blanco
MEP S&D, Parlamento Europeo

El borrador de la ley de Inteligencia Artificial del Parlamento está culminando y se espera su aprobación en el próximo pleno, lo que implicaría que el triálogo se realizará durante la presidencia española del Consejo de la Unión Europea en el segundo semestre de este año. El grupo S&D ha sido exigente con el tema de los usos prohibidos y ha insistido en la eliminación de excepciones como la vigilancia biométrica a distancia. También han puesto énfasis en la herramientas de concientización de la sociedad sobre las oportunidades y riesgos de la Inteligencia Artificial, así como en la gobernanza.

Indica que se ha avanzado en temas de modelos fundacionales y se espera la implementación de medidas importantes como exigir que se señale que se está interactuando con una IA. Hay consenso en que la regulación debe abarcar no solo temas sensibles, sino también el uso general de las aplicaciones.

Menciona la importancia de las consultas públicas y el momento propicio actual para hacer propuestas. En cuanto al sandbox español, se ha tenido contacto regular, pero no en profundidad, y se espera ver cómo avanza.

En relación a la implementación del reglamento, destaca la necesidad de negociar y ajustar las obligaciones, pero se considera que hay una buena regulación y no se anticipan problemas. Señala que la regulación es esperada a nivel mundial y se destaca la importancia de marcar la pauta en la materia.

En cuanto a la gobernanza a nivel europeo, menciona que ha habido evolución en las posiciones y se plantea la creación de una oficina desde el Parlamento que se asemeje a una agencia, pero sin ser nombrada como tal por cuestiones burocráticas y de presupuesto. Destaca la necesidad de una gran apuesta europea y de un instrumento político-administrativo potente.

Por último, considera inevitable que exista una regulación sobre el uso militar de la Inteligencia Artificial y menciona un informe del Parlamento Europeo al respecto en 2020, por lo que espera que la Comisión cumpla con la promesa de una regulación específica sobre este tema a futuro.

Beatriz García del Pozo
Responsable de Calidad, Normativas técnicas y Seguridad en INCIBE - Instituto Nacional de Ciberseguridad

García del Pozo menciona que el instituto es una entidad pública dependiente del Ministerio de Asuntos Económicos y Transformación Digital. Su misión principal es mejorar la ciberseguridad y confianza digital de los ciudadanos y empresas en España. También se enfoca en proteger y defender a los ciudadanos menores y potenciar la industria española en ciberseguridad, así como promover la investigación y desarrollo en este campo.

El INCIBE colabora con diferentes organizaciones a nivel nacional y gubernamental, y el desarrollo de la Estrategia Nacional de Inteligencia Artificial está liderado por la Secretaría de Estado de Digitalización e Inteligencia Artificial del Ministerio. La función del INCIBE en esta estrategia es asegurar que las tecnologías habilitadoras, incluyendo el 5G y la inteligencia artificial, cumplan con requisitos mínimos de ciberseguridad.

A nivel europeo, el instituto participa en grupos de trabajo de la Comisión Europea donde se están desarrollando estándares y certificaciones, especialmente en el ámbito de la certificación de sistemas y productos. Se espera que próximamente se publiquen los criterios de certificación para Cloud, pero se está a la espera de las directrices de regulación final por parte de la Comisión Europea.

La realización de ciberejercicios a nivel nacional e internacional se enfoca en el cumplimiento de la misión principal del INCIBE que es mejorar la ciberseguridad y confianza digital de los ciudadanos y empresas en España. En los ciberejercicios también se invita a participar a entidades públicas. El ámbito de la ciberseguridad en España está organizado a través de tres referentes nacionales y gubernamentales: el Centro Criptológico Nacional (CCN), que se encarga de las entidades públicas; el Comando Conjunto del Ciberespacio del Ministerio de Defensa, que se ocupa de las redes de defensa; y el Instituto Nacional de Ciberseguridad (INCIBE), que abarca a los ciudadanos menores y empresas privadas.

En estos momentos se está planeando desde la Secretaría de Estado de Digitalización e Inteligencia Artificial un centro de tecnologías en el ámbito de la ciberseguridad. Este centro cubrirá tecnologías como el 5G, Internet de las Cosas, sistemas de control industrial y diferentes aspectos de la inteligencia artificial. El objetivo es permitir a la industria nacional realizar pruebas en el campo de la ciberseguridad.

Lawrence Chan
Member of Technical Staff, Alignment Research Center (ARC)

Lawrence Chan menciona que en términos de evaluación de los sistemas de IA, ARC se ha enfocado en las capacidades peligrosas de los modelos y no en su alineación. Se tiene la ventaja de comenzar con modelos de lenguaje que han pasado por un proceso de entrenamiento y tienen predecesores sobre los cuales se espera que se puedan detectar signos de comportamiento problemático antes de que los modelos adquieran capacidades peligrosas.

Es más fácil obtener un comportamiento específico incentivando al modelo que esperar que se desarrolle espontáneamente. Sin embargo, se reconoce que existe la posibilidad de que los modelos muestren comportamientos engañosos y se discute la importancia de las técnicas de alineación y la verificación de sus suposiciones.

Los modelos actuales han experimentado un progreso incremental en términos de capacidades y todavía están lejos de alcanzar umbrales importantes. Destaca la importancia de evaluar la capacidad de replicación autónoma de los modelos y cómo esto puede ser relevante en términos de riesgo y pérdida de control; por otro lado, existen consideraciones importantes más allá de las capacidades peligrosas, como la no discriminación, la imparcialidad en general, evitar el lenguaje ofensivo, no ayudar en delitos, desinformación o spam, etc.

Chan plantea la idea de destinar más recursos a las evaluaciones de IA, argumentando que si se invierte tanto en la creación de modelos, también se debería estar dispuesto a invertir en evaluaciones de calidad. Menciona que las evaluaciones pueden ser significativamente más costosas que simplemente ejecutar conjuntos de datos estandarizados.

Destaca la importancia de la participación de los gobiernos en la mitigación de los riesgos de la IA. Indica que los gobiernos pueden requerir evaluaciones de seguridad, establecer estándares y promover la transparencia en el desarrollo de la IA, al mismo tiempo que existe una posibilidad de colaboración y coordinación entre los gobiernos de diferentes países,

especialmente aquellos en los que se concentran los principales laboratorios de IA y las empresas. En un escenario optimista, si los gobiernos pueden colaborar y compartir intereses comunes, pueden establecer regulaciones y estándares para la IA de manera efectiva y mitigar los riesgos asociados.

Sin embargo, concluye que los regímenes de evaluación y regulación no son una solución permanente y sostenible a largo plazo. Menciona que la historia ha demostrado que las regulaciones pueden no durar mucho tiempo y que los incentivos de los diferentes actores pueden cambiar con el tiempo. Estos regímenes de evaluación pueden servir como una transición gradual hacia un futuro con AGI, permitiendo más tiempo para la investigación en alineación y la obtención de consenso público y gubernamental sobre las acciones que se deben tomar.

Marius Hobbhahn
Director de Apollo Research

Apollo quiere hacer evaluaciones de modelos de IA basadas en capacidades, utilizando *prompting* y *fine-tuning* como técnicas principales. Consideran que las evaluaciones de capacidades están más desatendidas que las evaluaciones de alineamiento.

En este contexto, quieren enfocarse en el alineamiento engañoso por (1) ser una capacidad instrumental para llevar a cabo muchas de las acciones que pueden causar daño a gran escala y (2) ser una cuestión desatendida. La operacionalización de esta capacidad conlleva, entre otros, detectar si el modelo tiene conciencia situacional y tiene la motivación de mantener sus objetivos.

Apollo plantea hacer evaluaciones durante el entrenamiento por dos motivos: (1) algunas capacidades pueden surgir a partir de cierto cómputo de entrenamiento y (2) falta comprensión sobre cómo y por qué surgen estas capacidades. Hobbhahn argumenta que las evaluaciones durante el entrenamiento son baratas y no requieren pausar el entrenamiento. También cree que los modelos se están acercando al punto de desarrollar capacidades peligrosas, y que hay que empezar a establecer medidas estrictas de precaución –por ejemplo, asegurar que un sistema se puede apagar–.

Las evaluaciones durante el diseño también parecen deseables, pero hay menos claridad sobre cómo predecir capacidades. Algunas ideas son (1) utilizar leyes de escala basadas en cómputo y (2) hacer tests cualitativos para ver el grado de mejora del modelo entre dos puntos de referencia. Hobbhahn opina que los planes de entrenamiento de modelos sobre un determinado umbral deban ser aprobados por una autoridad.

Hobbhahn considera importante desarrollar ya estándares basados en principios generales para guiar los procesos de auditoría. Al mismo tiempo, estos procesos deben alimentar los estándares creados en un proceso iterativo para perfeccionar tanto estándares como auditorías.

Se discuten los riesgos asociados a la ejecución de auditorías. Hobbhahn considera oportuno no publicar la mayor parte de los resultados para evitar usos indebidos de esa

información. Sin embargo, aclara que para algunas evaluaciones tiene sentido publicar todo y para otras no, a lo que se debería pensar en qué caso se está antes de publicar.

También considera importante que las evaluaciones se lleven a cabo en entornos controlados para evitar filtraciones. Algunas ideas para ello incluyen (1) utilizar APIs, (2) hacer la evaluación en el hardware de la empresa o (3) diseñar la evaluación para que la empresa la ejecute en un proceso supervisable por la organización evaluadora.

Apéndice 2. Revisión de la literatura

Resumen de las secciones de la revisión de literatura para la cual se dividieron las fuentes en tres categorías: Investigación estratégica, investigación de políticas y documentos oficiales.

1. Investigación estratégica

La investigación estratégica en el área de inteligencia artificial ha cobrado gran relevancia en los últimos años debido a los beneficios que esta tecnología puede aportar a la sociedad, así como los riesgos y desafíos que implica su desarrollo y uso. En este sentido, diversas fuentes especializadas han abordado diferentes aspectos de este tema, aportando información y perspectivas valiosas para la comprensión y gestión de la IA. A continuación, se presentan algunas de estas fuentes:

En primer lugar, (European Commission. Directorate General for Communications Networks, Content and Technology. & Grupo de expertos de alto nivel sobre inteligencia artificial., 2019) (Scharre, 2019) destaca la importancia de **desarrollar directrices éticas para garantizar la confianza en la IA**, en línea con los **valores y derechos fundamentales** de la Unión Europea. Por su parte, el informe de (Brundage, Avin, Clark, Toner, Eckersley, Garfinkel, Dafoe, Scharre, Zeitzoff, Filar, et al., 2018) (Brundage et al., 2020) señala la necesidad de **prever y mitigar posibles usos malintencionados de la IA,** así como de **fomentar la colaboración internacional** para este fin.

En cuanto a los riesgos y beneficios de la IA, (Conn, 2015) destaca que, si bien esta tecnología puede aportar importantes avances en áreas como la salud, el transporte o la energía, también puede tener efectos negativos en la privacidad, el empleo o la seguridad. En este sentido, (Yudkowsky, 2008) advierte sobre el riesgo de la IA como factor de riesgo global, mientras que el informe de (Hatzius et al., 2023) destaca el potencial de la IA para impulsar el crecimiento económico., pero alerta con los peligros que existen para los empleos en diferentes sectores a nivel global, encontrando que en la **zona Euro son vulnerables hasta un 25% de las posiciones laborales.** mientras que la detección y mitigación de amenazas emergentes, como las operaciones de influencia automatizadas en el público general, es discutida por (Goldstein et al., 2023)

Por su parte, (Ngo, 2020) y (Amodei et al., 2016b) aborda la importancia de **alinear los objetivos de la IA con los intereses humanos** para evitar posibles consecuencias indeseadas. Además de proponer una serie de medidas como: **Cascos de seguridad o**

***sandbox*, control en el diseño, experimentación y pruebas rigurosas, transparencia y explicabilidad** en este último tema existen grandes incógnitas que incrementan la impredecibilidad de los sistemas. En línea con esto, el informe de (*AI Alignment 2018-19 Review - AI Alignment Forum*, 2020) destaca la necesidad de **investigar y desarrollar soluciones para garantizar la seguridad y la alineación de la IA.**

En pro de disminuir los riesgos asociados a las IA (Babcock et al., 2016) plantea una serie de parámetros como el **Diseño seguro**, Esto incluye la incorporación de mecanismos de seguridad y la minimización de las vulnerabilidades. **Verificación y validación** probar y validar la AGI antes de su implementación punto en el coincide con (Ngo, 2020) , **Monitoreo y control continuo** para tener tener la capacidad de controlarla en caso de que se vuelva peligrosa o inesperada y por último la **contención física** se refiere a la necesidad de tener medidas físicas en su lugar para contener la AGI en caso de que se vuelva incontrolable, como los "**Kill Switch".** El artículo (*AI Alignment Forum*, 2020), se describen 11 propuestas para construir IA avanzada y segura en las que destacan la creación de una red de **agencias reguladoras, la cooperación internacional** en materia de investigación, desarrollo, gestión del riesgo y formulación de políticas, **Anticipar riesgos desde las etapas de diseño, Incorporar mecanismos de "kill switch"**, **transparencia y explicabilidad**, **Verificabilidad** a través de la **auditoría y certificación** que cumpla con los **estándares internacionales**, **Establecer estándares de seguridad y ética** para los desarrolladores de IA, fomentar la **educación** y la conciencia pública sobre los riesgos y beneficios de la IA.

En relación con los avances en IA, el artículo de (Mnih et al., 2015) destaca el potencial de la IA para alcanzar niveles de control y aprendizaje comparables a los humanos mediante el aprendizaje por refuerzo profundo. En este sentido, la investigación de (Amodei & Hernandez, 2018), de OpenIA muestra el crecimiento exponencial de la potencia de cómputo utilizada en la IA en los últimos años. Por último, (Carlsmith, 2022) aborda el riesgo de la IA como agente de búsqueda de poder y control, lo que podría tener consecuencias existenciales para la humanidad. En este sentido, el informe de (Scharre, 2019) (Brundage, Avin, Clark, Toner, Eckersley, Garfinkel, Dafoe, Scharre, Zeitzoff, Filar, et al., 2018) destaca la importancia de **prever y regular el uso de "killer apps" basadas en la IA**, que podrían tener **efectos graves en la seguridad nacional y global**.

En conclusión, la investigación estratégica en IA implica abordar diferentes aspectos, desde la **ética y la confianza** hasta la **seguridad y la alineación con los intereses humanos**. Para ello, es necesario contar con una **perspectiva multidisciplinaria y colaborativa** que **tenga en cuenta tanto los beneficios como los riesgos de esta tecnología**, y que trabaje en la búsqueda de soluciones para **maximizar sus beneficios y minimizar sus riesgos y desafíos.**

2. Investigación de políticas

La inteligencia artificial (IA) es una tecnología en rápido desarrollo que ha generado interés en el ámbito político. Las implicaciones de la IA son vastas y, por lo tanto, se han llevado a cabo varias investigaciones políticas en el área. A continuación, se resumen algunas de las investigaciones más importantes que se han realizado en este campo:

El artículo de (Brundage et al., 2020) se centra en el desarrollo de IA confiable. La confianza en la IA es esencial para su aceptación en la sociedad. Este trabajo propone un mecanismo para apoyar las afirmaciones verificables en el desarrollo de IA. La verificación de la IA se puede hacer mediante la certificación, que se discute en el artículo de (Cihon et al., 2021). **La certificación es un proceso que puede ayudar a reducir las asimetrías de información en la práctica ética de la IA**. Propone un marco de certificación de la IA para mejorar la transparencia y la responsabilidad en su desarrollo y uso. El informe destaca la necesidad de una **evaluación independiente y estandarizada de los sistemas de IA**, y la importancia de la **divulgación de información sobre el desempeño y la seguridad de la IA.**

La seguridad es otra preocupación importante en el desarrollo de IA. Construir IA avanzada y segura se destacan los siguientes puntos: la creación de una red de **agencias reguladoras, la cooperación internacional** en materia de investigación, desarrollo, gestión del riesgo y formulación de políticas, **Anticipar riesgos desde las etapas de diseño, Incorporar mecanismos de "kill switch"**, **transparencia y explicabilidad**, **Verificabilidad** a través de la **auditoría y certificación** que cumpla con los **estándares internacionales**, **Establecer estándares de seguridad y ética** para los desarrolladores de IA, fomentar la **educación** y la conciencia pública sobre los riesgos y beneficios de la IA, puntos en los cuales coincide el informe de *"Future Proof"* del Centro para la Resiliencia a Largo Plazo (Ord et al., 2021), el informe *"AI Governance: A Research Agenda"* (Dafoe, 2018) y el informe *"Policymaking in the Pause"* (Future Of Life Institute, 2023).

La propuesta de "Auditing large language models: A three-layered approach" de (Mökander et al., 2023) aborda específicamente la s**eguridad de los modelos de lenguaje**. El artículo describe una forma en que los modelos de lenguaje se pueden auditar en tres capas: la **capa de entrada, la capa de atención y la capa de salida**. **Esta estrategia puede ayudar a identificar y prevenir la propagación de información errónea o potencialmente peligrosa.**

El papel de la cooperación en el desarrollo responsable de la IA es abordada por (Askell et al., 2019). La **cooperación** es necesaria para construir **IA responsable**, ya que la **responsabilidad en el desarrollo de IA es compartida** por muchas partes interesadas. La regulación de la IA es otro tema importante que se discute en el artículo de (European Commission., 2021). Este artículo examina el panorama de estandarización de la IA listando los estandares más importantes como: ISO/IEC 23894:2020, IEEE P7003, ISO/IEC 30141:2019, NIST SP 800-53, IEEE 1291 y cómo se relacionan con la propuesta de la Comisión Europea para un marco regulatorio de la IA, También se encuentra un estandar publicado este año ISO/IEC DIS 42001.

El artículo de (Whittlestone & Clark, 2021) sugiere que la IA tiene un gran potencial para afectar la sociedad, y los gobiernos tienen un papel importante en **garantizar que se desarrolle de manera responsable y justa**. En este sentido, se propone que los **gobiernos monitoreen el desarrollo de la IA**. Para ello, se hacen sugerencias en materia de La **gobernanza corporativa de la IA** también se discute en el artículo de (Cihon et al., 2021) el artículo propone que las empresas adopten prácticas de gobernanza corporativa para garantizar que la IA se desarrolle en interés público.

Por último, el artículo (Tucker et al., 2020) " aborda cómo la eficiencia de los datos puede afectar la sociedad y la gobernanza de la IA de forma positiva en la toma de mejores decisiones. Además, la mejora en la eficiencia de los datos puede ayudar a reducir costos y aumentar la productividad. El aumento de la eficiencia de los datos puede tener implicaciones negativas como el aumento de la vigilancia, la violación de la privacidad perdida de la libertad, también discute cómo puede **exacerbar las desigualdades existentes**. Por ejemplo, puede llevar a un **sesgo en los algoritmos de toma de decisiones**, lo que puede tener **consecuencias negativas** para los **grupos sociales marginados o subrepresentados**. Además, puede llevar a una **concentración de poder en manos de algunas grandes corporaciones** que controlan vastas cantidades de datos. En este caso los gobiernos deben tomar mayor acción para evitar estos escenarios.

El informe de (Siegmann & Anderljung, 2022) The Brussels Effect and Artificial Intelligence: How EU regulation will impact the global AI market discute el impacto del **Reglamento General de Protección de Datos (GDPR) de la UE en la regulación global de la inteligencia artificial (IA)**, destacando los desafíos planteados por la aplicación extraterritorial del GDPR y su **interacción con otros marcos regulatorios de IA**. Se argumenta que la próxima regulación de la IA en la UE está lista para tener un impacto global similar debido a la capacidad de la **UE para influir en la normativa global a través de su poder de mercado y sus estándares regulatorios.** La existencia de un efecto Bruselas significativo **puede conducir a una regulación más estricta de la IA a nivel mundial**, lo que aumenta la importancia de ayudar a dar forma al régimen regulatorio de la IA de la UE. Es esencial garantizar que la regulación de la IA de la UE esté preparada para el futuro y pueda adaptarse a un mundo de capacidades de IA cada vez más transformadoras.

3. Documentos oficiales

En los últimos años, varios gobiernos han presentado documentos y estrategias relacionados con la inteligencia artificial (IA). A continuación, se presenta un resumen de los puntos relevantes de seis documentos gubernamentales que abordan la IA.

- *A Pro-Innovation Approach to AI Regulation*, (Department for Science, Innovation & Technology, 2023) es un documento del gobierno británico que aborda la regulación de la IA. El documento aboga por un enfoque pro-innovación que permita a las empresas experimentar con la IA sin excesiva regulación. El gobierno reconoce que la IA tiene el potencial de mejorar muchos aspectos de la vida, pero también reconoce la necesidad de establecer ciertas regulaciones para garantizar que la IA sea segura y confiable.

- La Comisión Europea presentó en 2021 la "Artificial Intelligence Act", que establece reglas armonizadas para la IA en toda la Unión Europea. La legislación clasifica la IA en cuatro categorías: IA inaceptable, IA de alto riesgo, IA de riesgo limitado y IA de uso permitido. Las IA de alto riesgo estarán sujetas a requisitos de seguridad y transparencia más estrictos.

- La Estrategia Nacional de Inteligencia Artificial de España (SEDIA, 2020) (vigente), presentada en 2020, establece la necesidad de invertir en investigación y desarrollo

en IA para mantener a España a la vanguardia de la tecnología. También se destacó la importancia de la colaboración público-privada y la necesidad de garantizar que la IA se utilice de manera ética y responsable.

- La Estrategia Española para la Inteligencia Artificial en I+D+I. (Ministerio de Ciencia, Innovación, 2019) (complementario o no vigente) se centra en la aplicación de la IA en la investigación médica y la salud. La estrategia establece objetivos específicos, como desarrollar algoritmos de IA para mejorar la detección temprana de enfermedades y mejorar la eficiencia de los ensayos clínicos.

- El Plan de Recuperación, Transformación y Resiliencia de España (Gobierno de España, 2021b), incluye menciones en tres ejes específicos uno de ellos Transformación digital plasmando las inversiones significativas en tecnologías digitales, incluida la IA que se realizarán hasta el 2025. El plan establece que la inversión en tecnología digital es fundamental para la recuperación económica y la transformación de la economía española.

- "España Digital 2026" (Gobierno de España, 2021a) es una estrategia del gobierno español que establece una hoja de ruta para la transformación digital de España en los próximos cinco años. La estrategia destaca la importancia de la IA para la economía española y establece objetivos específicos para su desarrollo, como la creación de una estrategia nacional de IA y la inversión en proyectos de investigación y desarrollo.

Los documentos gubernamentales analizados destacan la **importancia de la IA para la economía y el bienestar social. Los gobiernos reconocen la necesidad de establecer regulaciones** para garantizar que la IA sea **segura y confiable,** al tiempo que **fomentan la innovación y el desarrollo en el campo de la IA**. También se destaca la importancia de la **colaboración público-privada** y la **inversión en investigación y desarrollo** para mantenerse a la vanguardia de la tecnología.

Apéndice 3. Mapeo de actores

Entender el desarrollo del *sandbox* implica también comprender la estructura de actores que se está desarrollando en Europa sobre este tema y la asignación de funciones que se están dando en el contexto de la gobernanza y la regulación. Para conocer esta, se llevaron a cabo una serie de entrevistas con expertos españoles e internacionales, así como con actores institucionales relacionados con el desarrollo del *sandbox*.

Las personas entrevistadas fueron seleccionadas a partir del contenido del informe y la disponibilidad de los mismos, dada la agitación que existe a nivel mundial en relación con el desarrollo de la IA.

El contenido del informe está pensado para abordar la inteligencia artificial desde la oportunidad histórica de la publicación del Reglamento de la UE para la IA como la primera ley de este tema que abarca 27 países en total (y que otros países como EE.UU aún no la

desarrollan), y cómo esta oportunidad histórica se desarrolla a partir de unos riesgos que se derivan del uso continuo y la expansión del conocimiento de inteligencia artificial que son aprovechados por una lista de vectores que toman ventaja de las vulnerabilidades de los sistemas que se convierten en amenazas concretas, así como existen una serie de riesgos estructurales que transforman a gran escala los sistemas culturales, sociales y económicos del mundo entero; para luego terminar analizando la propuesta de Reglamento europeo de cara a estos vectores, amenazas y riesgos, de cómo el *sandbox* se constituye en una herramienta para probar propuestas, regulaciones y tecnología de cara a permitir el desarrollo de todo un mercado y una industria de forma segura para la humanidad.

Con base en esta estructura se escogieron una serie de expertos y actores institucionales que ayudaron a la comprensión de cada uno de los temas.

En cuanto a los expertos se contó con la participación de tres de nacionalidad española, José H. Orallo, Toni Lorente y Pablo Villalobos, y de otros cuatro de distintas nacionalidades, Charlotte Siegmann, Markus Anderljung, Samuel Hilton y Riso Uuk; los expertos españoles permitieron detallar el estado actual de la IA en el contexto global y local, así como los desafíos a los que el reglamento europeo se enfrenta de cara al desarrollo del *sandbox* y el estado acual de la industria local en cuanto a tecnología e inteligencia artificial; por el lado de los expertos globales, en particular Charlotte Siegmann y Markus Anderljung fueron de gran ayuda para comprender el Reglamento europeo, su alcance y, en específico, las pruebas que este tendrá que enfrentar de cara a fomentar la gobernanza y al regulación de los sistemas de IA, y cómo esto tendrá un efecto a nivel global si se realiza correctamente; por el lado de las comunicaciones, Samuel Hilton permitió entender los retos que los diferentes actores (entre gobierno, privados y académicos) tendrán al momento de poder hacer efectivas las normas y recomendaciones que surjan en el desarrollo de las diferentes herramientas políticas, jurídicas y económicas que integran la IA; por último, pero con especial agrado, Risto Uuk nos ayudó a comprender los alcances de la IA a nivel global y en específico, de cómo el *sandbox* será una herramienta útil para probar instituciones y capacidades, y de cómo es un gran reto del cual podrán derivarse lecciones útiles que llevan a una comprensión de la dimensión de lo que significa el desarrollo de los sistemas de IA para la humanidad.

En cuanto a lo institucional, el acceso a información ha sido más limitado bajo el contexto que el reglamento se encuentra a la espera de ser discutido por los tres órganos representativos de la Unión Europea y que el *sandbox* no ha comenzado formalmente. Todavía existe un grado de incertidumbre de responsabilidades y funciones dentro del *sandbox* que hace que las instituciones que hasta el momento trabajan al respecto (como la Secretaría de Estado de Digitalización e Inteligencia Artificial, el Consejo Asesor de Inteligencia Artificial, el Instituto Nacional de Ciberseguridad, entre otros) y las no creadas todavía (como la Agencia Española de Inteligencia Artificial), se encuentren a la espera de definir más concretamente cómo funcionará el *sandbox* y qué responsabilidades asumirá cada una en este.

Apéndice 4. Reglamento europeo para la Inteligencia Artificial

El principal punto de partida de la propuesta de ley es la clasificación de los sistemas de IA según el nivel de riesgo que implican. En concreto, la propuesta se basa en una jerarquía que distingue entre riesgos inaceptables, altos, limitados y mínimos. Los dos primeros son el objeto principal de la regulación.

Como parte de la categoría de **riesgos inaceptables**, quedan prohibidas prácticas que supongan una amenaza clara para la seguridad, los medios de subsistencia y los derechos de las personas. Por el momento, tres prácticas se han considerado inaceptables por ir en contra de los valores europeos: alterar el comportamiento humano para causar daño; evaluar y clasificar personas según su conducta social; y usar sistemas de identificación biométrica remota en tiempo real en espacios públicos, salvo en casos de emergencia.

Por otro lado, los sistemas de **alto riesgo** son aquellos con el potencial de causar un mayor impacto por desplegarse en sectores críticos, incluyendo las infraestructuras esenciales, la educación, el empleo, los servicios públicos y privados esenciales, la aplicación de la ley y la gestión de fronteras. En este caso, varios requerimientos pesan sobre el desarrollo e implementación de todos los productos.

En primer lugar, se requiere que los proveedores de sistemas de alto riesgo establezcan, implementen, documenten y mantengan un **sistema de gestión de riesgos** en dos fases. Primeramente, se deberán identificar y evaluar riesgos conocidos y previsibles, tanto antes como después de la comercialización. Los riesgos podrán considerarse “conocidos” o “previsibles” si el desarrollador del sistema de IA debería conocerlo al adoptar un nivel razonable de diligencia. Actualmente, pero, el Reglamento no explica claramente qué constituye “un nivel razonable de diligencia”.

La segunda fase consiste en reducir los riesgos detectados a un nivel aceptable: los proveedores deberán eliminar completamente los riesgos en la medida de lo posible o, en caso contrario, implementar medidas de mitigación y control, así como entrenar a los usuarios para que hagan un uso responsable. De esta manera, el sistema de gestión de riesgos será un proceso a repetir hasta que todos los riesgos identificados sean aceptables. La identificación de riesgos inaceptables que no puedan ser reducidos implicará la detención inmediata del desarrollo y/o despliegue del sistema de IA en cuestión (Schuett, 2023b).

En paralelo, los proveedores desarrollarán un **sistema de gestión de calidad** que asegure que el desarrollo y verificación del sistema de IA cumplan con el Reglamento. Antes de la salida al mercado, los desarrolladores deberán proporcionar **documentación técnica** que incluya detalles del diseño y la arquitectura del sistema. Adicionalmente, los **conjuntos de datos de entrenamiento** deberán haber seguido pautas de gobernanza referentes a la elección de un diseño adecuado, operaciones de tratamiento oportunas y detección de posibles deficiencias y sesgos.

También se expondrán los procedimientos destinados a reforzar la **ciberseguridad** y la **robustez**, es decir, la resistencia del sistema ante alteraciones. Paralelamente, se requerirán medidas de **transparencia** como facilitar instrucciones de uso accesibles y, cuando aplique, informar al usuario de que está interactuando con una IA. Con base en la documentación, se realizarán **procedimientos de evaluación** mayoritariamente internos.

En caso de pasar esta examinación, los sistemas de IA quedarán avalados por una declaración de conformidad redactada por el proveedor y a disposición de las autoridades.

Durante todo el periodo de uso, los sistemas deberán estar **supervisados por humanos** que comprendan las capacidades y limitaciones del modelo y puedan intervenir en su funcionamiento en caso de ser necesario. En paralelo, los eventos (*logs*) ocurridos durante todo el ciclo de vida serán **registrados automáticamente** para garantizar trazabilidad. En el seguimiento posterior a la comercialización, todo incidente grave o fallo deberá ser notificado. En este caso, las autoridades europeas de **vigilancia del mercado** tendrán derecho a acceder a datos, documentación y código fuente. Cuando el operador sea incapaz de adoptar medidas correctoras, también se les concederá la potestad de prohibir o restringir la comercialización del sistema.

Para la implementación de la regulación, la UE apuesta por la creación de entornos controlados de prueba o ***sandboxes***, los cuales tienen el objetivo de identificar y solucionar problemas potenciales en la aplicación del Reglamento. Estos entornos estarán disponibles mediante convocatoria, para que empresas y organizaciones que deseen probar nuevas soluciones de IA participen en ellos. Los proyectos seleccionados para integrar los *sandboxes* podrán compartir información y conocimientos, fomentando así la colaboración y el intercambio de experiencias y mejores prácticas. Además, recibirán acceso a asesoramiento y orientación de expertos, contando con un entorno seguro y controlado para probar soluciones de IA antes de su lanzamiento al mercado. Los resultados de las pruebas realizadas contribuirán a los esfuerzos de la Comisión Europea en la aplicación eficaz del nuevo Reglamento y facilitarán la flexibilización y adaptación de las normas a las necesidades reales que demande esta tecnología.

En este contexto, la regulación ordena que se asignen autoridades nacionales de supervisión e introduce el Comité Europeo de Inteligencia Artificial como nexo de todos los organismos estatales. Durante el *sandbox*, las autoridades nacionales deberán presentar informes anuales al Comité y a la Comisión, incluyendo resultados, enseñanzas y recomendaciones.

Referencias

Acemoglu, D. (2021). *Harms of AI* (N.º w29247; p. w29247). National Bureau of Economic Research. https://doi.org/10.3386/w29247

Agudo, U., & Matute, H. (2021). The influence of algorithms on political and dating decisions. *PLOS ONE*, *16*(4), e0249454. https://doi.org/10.1371/journal.pone.0249454

Aguilar, X., & Markidis, S. (2021). *A Deep Learning-Based Particle-in-Cell Method for Plasma Simulations* (arXiv:2107.02232). arXiv. http://arxiv.org/abs/2107.02232

Aksela, M., Marchal, S., Patel, A., & Rosenstedt, L. (2022). *The security threat of AI-enabled cyberattacks* (p. 30). Finnish Transport and Communications Agency Traficom. https://www.traficom.fi/sites/default/files/media/publication/TRAFICOM_The_security_threat_of_AI-enabled_cyberattacks%202022-12-12_en_web.pdf

AMETIC. (2023). *MAPEO DEL ECOSISTEMA ESPAÑOL DE MICROELECTRÓNICA*. AMETIC. https://ametic.es/wp-content/uploads/2023/04/Mapeo_202304024_B.pdf

Amodei, D., & Hernandez, D. (2018). *AI and compute*. https://openai.com/research/ai-and-compute

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016a). *Concrete Problems in AI Safety* (arXiv:1606.06565). arXiv. http://arxiv.org/abs/1606.06565

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016b). *Concrete Problems in AI Safety* (arXiv:1606.06565). arXiv. http://arxiv.org/abs/1606.06565

Anderson, H. S., Woodbridge, J., & Filar, B. (2016). DeepDGA: Adversarially-tuned domain generation and detection. *Proceedings of the 2016 ACM workshop on artificial intelligence and security*, 13-21.

Anderson, J. (2021, abril 15). *The dynamics of data accumulation*. Bruegel | The Brussels-Based Economic Think Tank.

https://www.bruegel.org/blog-post/dynamics-data-accumulation

Askell, A., Brundage, M., & Hadfield, G. (2019). *The Role of Cooperation in Responsible AI Development*. https://doi.org/10.48550/ARXIV.1907.04534

Babcock, J., Kramar, J., & Yampolskiy, R. (2016). *The AGI Containment Problem*. https://doi.org/10.48550/ARXIV.1604.00545

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2020). *Emergent Tool Use From Multi-Agent Autocurricula* (arXiv:1909.07528). arXiv. http://arxiv.org/abs/1909.07528

Baker, M. (2023). *Nuclear Arms Control Verification and Lessons for AI Treaties* (arXiv:2304.04123). arXiv. http://arxiv.org/abs/2304.04123

Baum, S. D. (2020). Social choice ethics in artificial intelligence. *AI & SOCIETY*, *35*(1), 165-176. https://doi.org/10.1007/s00146-017-0760-1

Bawden, D., & Robinson, L. (2020). Information Overload: An Introduction. En D. Bawden & L. Robinson, *Oxford Research Encyclopedia of Politics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.1360

Bird, J., & Layzell, P. (2002). The evolved radio and its implications for modelling the evolution of novel sensors. *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)*, *2*, 1836-1841. https://doi.org/10.1109/CEC.2002.1004522

Blattner, L., & Nelson, S. (2021). How costly is noise? Data and disparities in consumer credit. *arXiv preprint arXiv:2105.07554*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. http://arxiv.org/abs/2108.07258

Bostrom, N., Dafoe, A., & Flynn, C. (2018). *Public Policy and Superintelligent AI: A Vector Field Approach*. https://nickbostrom.com/papers/aipolicy.pdf

Bou, C. P. (2023, marzo 9). *Los responsables del ciberataque al Hospital Clínic exigen un rescate a la Generalitat*. elperiodico. https://www.elperiodico.com/es/sanidad/20230309/ciberataque-ransom-house-hospital-clinic-rescate-ciberseguridad-84388678

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33), eabe5641. https://doi.org/10.1126/sciadv.abe5641

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., & Filar, B. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., … Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. https://doi.org/10.48550/ARXIV.1802.07228

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., … Anderljung, M. (2020). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. https://doi.org/10.48550/ARXIV.2004.07213

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*, 77-91.

Carlsmith, J. (2022). *Is Power-Seeking AI an Existential Risk?* https://doi.org/10.48550/ARXIV.2206.13353

Castillo, C. del. (2022, junio 29). *España es el país con más ciberataques para robar contraseñas o datos bancarios*. elDiario.es. https://www.eldiario.es/tecnologia/espana-pais-ciberataques-robar-contrasenas-datos

-bancarios_1_9129484.html

Chartoff, philip. (2018). *Perils of Lethal Autonomous Weapons Systems Proliferation: Preventing Non-State Acquisition | GCSP*. https://www.gcsp.ch/publications/perils-lethal-autonomous-weapons-systems-proliferation-preventing-non-state

Christiano, P. (2019). *What failure looks like*. https://www.lesswrong.com/posts/HBxe6wdjxK239zajf/what-failure-looks-like

Cihon, P., Kleinaltenkamp, M. J., Schuett, J., & Baum, S. D. (2021). AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries. *IEEE Transactions on Technology and Society*, *2*(4), 200-209. https://doi.org/10.1109/TTS.2021.3077595

Clarke, S., Whittlestone, J., Maas, M., Belfield, H., Hernández-Orallo, J., & Heigeartaigh, S. Ó. (2021). *Submission of Feedback to the European Commission's Proposal for a Regulation laying down harmonised rules on artificial intelligence* [Feedback]. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665626_en

Conn, A. (2015, noviembre 14). Benefits & Risks of Artificial Intelligence. *Future of Life Institute*. https://futureoflife.org/ai/benefits-risks-of-artificial-intelligence/

Consejo. (2022). *Artificial Intelligence Act: Council calls for promoting safe AI that respects fundamental rights*. https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/

Dafoe, A. (2018). *AI Governance: A Research Agenda* (p. 54). Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de las Casas, D., Donner, C., Fritz, L., Galperti, C., Huber, A., Keeling, J., Tsimpoukelli, M., Kay, J., Merle, A., Moret, J.-M., … Riedmiller, M. (2022). Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*,

602(7897), 414-419. https://doi.org/10.1038/s41586-021-04301-9

Department for Science, Innovation & Technology. (2023). *A pro-innovation approach to AI regulation, presented to Parliament by the Secretary of State for Science, Innovation and Technology by command of His Majesty*. Department for Science, Innovation & Technology.

Dezfouli, A., Nock, R., & Dayan, P. (2020). Adversarial vulnerabilities of human decision-making. *Proceedings of the National Academy of Sciences*, *117*(46), 29221-29228. https://doi.org/10.1073/pnas.2016921117

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1), eaao5580.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models* (arXiv:2303.10130). arXiv. http://arxiv.org/abs/2303.10130

Epoch. (2022). *Parameter, Compute and Data Trends in Machine Learning*. https://epochai.org/mlinputs/visualization

Erdil, E., & Besiroglu, T. (2022). Algorithmic progress in computer vision. *arXiv preprint arXiv:2212.05153*.

European Commission. Directorate General for Communications Networks, Content and Technology. & Grupo de expertos de alto nivel sobre inteligencia artificial. (2019). *Directrices éticas para una IA fiable.* Publications Office. https://data.europa.eu/doi/10.2759/14078

European Commission. Directorate General for Competition. (2019). *Competition policy for the digital era.* Publications Office. https://data.europa.eu/doi/10.2763/407537

European Commission. Joint Research Centre. (2021). *AI Watch, AI standardisation landscape state of play and link to the EC proposal for an AI regulatory framework.* Publications Office. https://data.europa.eu/doi/10.2760/376602

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). *Robust Physical-World Attacks on Deep Learning Models*

(arXiv:1707.08945). arXiv. http://arxiv.org/abs/1707.08945

Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., & Wen, Y. (2022). CapAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4064091

Fruhlinger, J. (2022, agosto 31). *Stuxnet explained: The first known cyberweapon*. CSO Online. https://www.csoonline.com/article/3218104/stuxnet-explained-the-first-known-cyberweapon.html

Future Of Life Institute. (2021). *FLI position on the Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Inteligencia-artificial-requisitos-legales-y-eticos/F2665546_es

Future Of Life Institute. (2023). *Policymaking in the Pause «What can policymakers do now to combat risks from advanced AI systems?»* (p. 14). FUTURE OF LIFE INSTITUTE. https://futureoflife.org/wp-content/uploads/2023/04/FLI_Policymaking_In_The_Pause.pdf

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., … Clark, J. (2022). *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*. https://doi.org/10.48550/ARXIV.2209.07858

GAWER, A., MANNE, G., STUCKE, M., VARIAN, H., & BURNSIDE, A. J. (2016). *Big data: Bringing competition policy to the digital era*. https://www.oecd.org/competition/big-data-bringing-competition-policy-to-the-digital-era.htm

Gobierno de España. (2021a). *España Digital 2026* (p. 159). Gobierno de España, Unión Europea.

https://espanadigital.gob.es/sites/espanadigital/files/2022-07/Espa%C3%B1aDigital_2026.pdf

Gobierno de España. (2021b). *Plan de Recuperación, Transformación y Resiliencia* (p. 348). Gobierno de España. https://www.lamoncloa.gob.es/temas/fondos-recuperacion/Documents/160621-Plan_Recuperacion_Transformacion_Resiliencia.pdf

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations* (arXiv:2301.04246). arXiv. http://arxiv.org/abs/2301.04246

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Hall, M., van der Maaten, L., Gustafson, L., Jones, M., & Adcock, A. (2022). A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*.

Hatzius, J., Briggs, joseph, Kodnani, D., & Pierdomenico, G. (2023). The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani). *Goldman Sachs Economic Research*. Goldman Sachs. https://www.key4biz.it/wp-content/uploads/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf

Hazell, J. (2023). Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns. *arXiv preprint arXiv:2305.06972*.

Hubinger, E. (2020). *An overview of 11 proposals for building safe advanced AI - AI Alignment Forum*. https://www.alignmentforum.org/posts/fRsjBseRuvRhMPPE5/an-overview-of-11-proposals-for-building-safe-advanced-ai

Hwang, T. (2018). *Computational Power and the Social Impact of Artificial Intelligence* (arXiv:1803.08971). arXiv. http://arxiv.org/abs/1803.08971

INCIBE. (2023). *Red Team Aguas Misteriosas | INCIBE-CERT | INCIBE*.

https://www.incibe.es/incibe-cert/blog/red-team-aguas-misteriosas

Is Retraining Displaced Workers a Good Investment? - Federal Reserve Bank of Chicago. (s. f.). Recuperado 16 de mayo de 2023, de https://www.chicagofed.org/publications/economic-perspectives/2005/2q-jacobson-lalonde-sullivan

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Dai, W., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1-38. https://doi.org/10.1145/3571730

Jiménez, M. (2022, octubre 17). *Adigital lanzará en enero un certificado de transparencia algorítmica para empresas*. Cinco Días. https://cincodias.elpais.com/cincodias/2022/10/16/companias/1665910625_275850.html

Jiménez, P. (2023). *What to expect from Europe's first AI oversight agency*. AlgorithmWatch. https://algorithmwatch.org/en/what-to-expect-from-europes-first-ai-oversight-agency/

Kates-Harbeck, J., Svyatkovskiy, A., & Tang, W. (2019). Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, *568*(7753), 526-531. https://doi.org/10.1038/s41586-019-1116-4

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology*, *24*(3), 36. https://doi.org/10.1007/s10676-022-09643-0

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020). *Specification gaming: The flip side of AI ingenuity*. https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity

KREPS, S. (2021). *DEMOCRATIZING HARM: ARTIFICIAL INTELLIGENCE IN THE HANDS OF NONSTATE ACTORS*. Brookings Institution. https://www.brookings.edu/wp-content/uploads/2021/11/FP_20211122_ai_nonstate_actors_kreps

Lamata, M. T., Pelta, D. A., & Verdegay, J. L. (2021). The Role of the Context in Decision and Optimization Problems. En M.-J. Lesot & C. Marsala (Eds.), *Fuzzy Approaches*

for Soft Computing and Approximate Reasoning: Theories and Applications (Vol. 394, pp. 75-84). Springer International Publishing. https://doi.org/10.1007/978-3-030-54341-9_7

Li, J., & Chignell, M. (2022). FMEA-AI: AI fairness impact assessment using failure mode and effects analysis. *AI and Ethics*, *2*(4), 837-850. https://doi.org/10.1007/s43681-022-00145-9

Longpre, S., Storm, M., & Shah, R. (2022). Lethal autonomous weapons systems & artificialintelligence: Trends, challenges, and policies. *MIT Science Policy Review*, *3*, 47-56. https://doi.org/10.38105/spr.360apm5typ

McGregor, S., Paeth, K., & Lam, K. (2022). *Indexing AI Risks with Incidents, Issues, and Variants* (arXiv:2211.10384). arXiv. http://arxiv.org/abs/2211.10384

Miller, C. (2021, noviembre 11). *Throwback Attack: BlackEnergy attacks the Ukrainian power grid*. Industrial Cybersecurity Pulse. https://www.industrialcybersecuritypulse.com/threats-vulnerabilities/throwback-attack-blackenergy-attacks-the-ukrainian-power-grid/

Ministerio de Asuntos Económicos y Transformación Digital. (2022, junio 26). *El Gobierno de España presenta, en colaboración con la Comisión Europea, el primer piloto del sandbox de regulación de Inteligencia Artificial en la UE*. Mineco. https://portal.mineco.gob.es/es-es/comunicacion/Paginas/20220627-PR_AI_Sandbox.aspx

Ministerio de Ciencia, Innovación. (2019). *ESTRATEGIA ESPAÑOLA DE I+D+I EN INTELIGENCIA ARTIFICIAL* (p. 48). Ministerio de Ciencia, Innovación y Universidades,. https://www.ciencia.gob.es/dam/jcr:5af98ba2-166c-4e63-9380-4f3f68db198e/Estrategia_Inteligencia_Artificial_IDI.pdf

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015).

Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529-533. https://doi.org/10.1038/nature14236

Mokander, J., Axente, M., Casolari, F., & Floridi, L. (2022). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines*, *32*(2), 241-268. https://doi.org/10.1007/s11023-021-09577-4

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). *Auditing large language models: A three-layered approach*. https://doi.org/10.48550/ARXIV.2302.08500

Munn, L. (2020). Angry by design: Toxic communication and technical architectures. *Humanities and Social Sciences Communications*, *7*(1), 53. https://doi.org/10.1057/s41599-020-00550-7

Nagesh, A. (2017, septiembre 18). Stanislav Petrov—The man who quietly saved the world—Has died aged 77. *Metro*. https://metro.co.uk/2017/09/18/stanislav-petrov-the-man-who-quietly-saved-the-world-has-died-aged-77-6937015/

Ngo, R. (2020). *AGI Safety From First Principles*. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ

Ngo, R. (2022). The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

Ngo, R., Chan, L., & Mindermann, S. (2022). *The alignment problem from a deep learning perspective*. https://doi.org/10.48550/ARXIV.2209.00626

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022). Deep Learning for Deepfakes Creation and Detection: A Survey. *Computer Vision and Image Understanding*, *223*, 103525. https://doi.org/10.1016/j.cviu.2022.103525

OECD. (2023). *Moving forward on data free flow with trust: New evidence and analysis of business experiences* (OECD Digital Economy Papers N.º 353; OECD Digital Economy Papers, Vol. 353). https://doi.org/10.1787/1afab147-en

O'Keefe, C., Cihon, P., Flynn, C., Garfinkel, B., Leung, J., & Dafoe, A. (2020). *The Windfall Clause: Distributing the Benefits of AI.* (p. 66). Centre for the Governance of AI Research Report. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/windfallclause/

Omohundro, S. M. (2007). The nature of self-improving artificial intelligence. *Singularity Summit*, *2008*.

OpenAI. (2023). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. http://arxiv.org/abs/2303.08774

Ord, T., Mercer, A., & Dannreuther, S. (2021). *FUTURE PROOF: THE OPPORTUNITY TO TRANSFORM THE UK'S RESILIENCE TO EXTREME RISKS* (p. 52). The Centre for the Study of Existential Risk. https://www.cser.ac.uk/media/uploads/files/Future_Proof_report_June_2021.pdf

Owen, D. (2023). *Extrapolating performance in language modeling benchmarks*. https://epochai.org/blog/extrapolating-performance-in-language-modelling-benchmarks

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). *Red Teaming Language Models with Language Models* (arXiv:2202.03286). arXiv. http://arxiv.org/abs/2202.03286

Perez, F., & Ribeiro, I. (2022). *Ignore Previous Prompt: Attack Techniques For Language Models* (arXiv:2211.09527). arXiv. http://arxiv.org/abs/2211.09527

Rabinowitch, E. (1961). Space Exploration in the Service of Science. *Bulletin of the Atomic Scientists*, *17*(5-6), 170-171. https://doi.org/10.1080/00963402.1961.11454217

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33-44.

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*,

118(26), e2024292118. https://doi.org/10.1073/pnas.2024292118

Rodríguez, B. (2022, septiembre 29). España Lanza el primer piloto de “AI regulatory sandbox” de la Unión Europea | Observatorio IA. *Observatorio IA de AMETIC*. https://observatorio-ametic.ai/regulacion-de-la-inteligencia-artificial/espana-lanza-el-primer-piloto-de-ai-regulatory-sandbox-de

Ruiz de Querol, R. (2022). *NO ES INEVITABLE: UN ALEGATO POR FUTUROS DIGITALES ALTERNATIVOS* (1.ª ed.). Alternativas económicas.

Russell, S. J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial intelligence: A modern approach* (3rd ed). Prentice Hall.

Sadeghi, M., & Arvanitis, L. (2023). Rise of the Newsbots: AI-Generated News Websites Proliferating Online. *NewsGuard*. https://www.newsguardtech.com/special-reports/newsbots-ai-generated-news-websites-proliferating

Scharre, P. (2018, septiembre 12). A Million Mistakes a Second. *Foreign Policy*. https://foreignpolicy.com/2018/09/12/a-million-mistakes-a-second-future-of-war/

Scharre, P. (2019, abril 16). Killer Apps. *Foreign Affairs*, *May/June 2019*. https://www.foreignaffairs.com/articles/2019-04-16/killer-apps?check_logged_in=1&utm_medium=promo_email&utm_source=lo_flows&utm_campaign=registered_user_welcome&utm_term=email_1&utm_content=20230404

Schippers, B. (2020). Artificial intelligence and democratic politics. *Political Insight*, *11*(1), 32-35.

Schuett, J. (2023a). *AGI labs need an internal audit function* (arXiv:2305.17038). arXiv. http://arxiv.org/abs/2305.17038

Schuett, J. (2023b). Risk Management in the Artificial Intelligence Act. *European Journal of Risk Regulation*, 1-19. https://doi.org/10.1017/err.2023.1

Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J. P., & Goldstein, T. (2021). *Just*

How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks (arXiv:2006.12557). arXiv. http://arxiv.org/abs/2006.12557

SEDIA. (2020). *ESTRATEGIA NACIONAL DE INTELIGENCIA ARTIFICIAL* (p. 90). Secretaría de Estado de Digitalización e Inteligencia Artificial adscrito a la Vicepresidencia Tercera del Ministerio de Asuntos Económicos y Transformación Digital. https://www.lamoncloa.gob.es/presidente/actividades/Documents/2020/ENIA2B.pdf

Seger, elizabeth, Avin, S., Pearson, G., Briers, M., Heigeartaigh, S. Ó., & Bacon, H. (2020). *Tackling threats to informed decisionmaking in democratic societies Promoting epistemic security in a technologically-advanced world* (p. 112). The Alan Turing Institute, Centre for the Study of Existential Risk, University of Cambridge. https://www.turing.ac.uk/sites/default/files/2020-10/epistemic-security-report_final.pdf

SETELECO. (2022). *Plan Sectorial Vigilancia Mercado Teleco 2022-2026*. SETELECO, Gobierno de España. https://avancedigital.mineco.gob.es/equipos-telecomunicacion/Documents/Plan-Sectorial-Vigilancia-Mercado-Teleco-2022-2026-1.pdf

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). *Compute trends across three eras of machine learning*. 1-8.

shah, R. (2020). *AI Alignment 2018-19 Review—AI Alignment Forum*. https://www.alignmentforum.org/posts/dKxX76SCfCvceJXHv/ai-alignment-2018-19-review

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., & Kenton, Z. (2022). *Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals* (arXiv:2210.01790). arXiv. http://arxiv.org/abs/2210.01790

Shavit, Y. (2023). *What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring* (arXiv:2303.11341). arXiv. http://arxiv.org/abs/2303.11341

Shevlane, T. (2022). *Structured access: An emerging paradigm for safe AI deployment*.

https://doi.org/10.48550/ARXIV.2201.05159

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., … Dafoe, A. (2023). *Model evaluation for extreme risks* (arXiv:2305.15324). arXiv. http://arxiv.org/abs/2305.15324

Shoshitaishvili, Y., Bianchi, A., Borgolte, K., Cama, A., Corbetta, J., Disperati, F., Dutcher, A., Grosen, J., Grosen, P., Machiry, A., & others. (2018). Mechanical phish: Resilient autonomous hacking. *IEEE Security & Privacy*, *16*(2), 12-22.

Siegmann, C., & Anderljung, M. (2022). *The Brussels Effect and Artificial Intelligence* (p. 97). Centre for the Governance of AI.

Song, Y., Zou, X., Gong, X., Becoulet, A., Buttery, R., Bonoli, P., Hoang, T., Maingi, R., Qian, J., Zhong, X., Liu, A., Li, E., Ding, R., Huang, J., Zang, Q., Liu, H., Wang, L., Zhang, L., Li, G., … EAST Team. (2023). Realization of thousand-second improved confinement plasma with Super I-mode in Tokamak EAST. *Science Advances*, *9*(1), eabq5273. https://doi.org/10.1126/sciadv.abq5273

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62-77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Temple, J. (2023). *This startup says its first fusion plant is five years away. Experts doubt it.* | *MIT Technology Review*. https://www.technologyreview.com/2023/05/10/1072812/this-startup-says-its-first-fusion-plant-is-five-years-away-experts-doubt-it/

Tucker, A. D., Anderljung, M., & Dafoe, A. (2020). Social and Governance Implications of Improved Data Efficiency. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 378-384. https://doi.org/10.1145/3375627.3375863

UNE. (2022). *Informe 2022 Vigilancia de Mercado*. UNE Observartorio vigilancia de mercado. https://www.une.org/normalizacion_documentos/Informe_anual_2022_OVM.pdf

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, *4*(3), 189-191. https://doi.org/10.1038/s42256-022-00465-9

Vincent, J. (2023, marzo 8). *Meta's powerful AI language model has leaked online—What happens now?* The Verge. https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse

Whittlestone, J., & Clark, J. (2021). *Why and How Governments Should Monitor AI Development* (arXiv:2108.12427). arXiv. http://arxiv.org/abs/2108.12427

Yudkowsky, E. (2008). *Artificial Intelligence as a Positive and Negative Factor in Global Risk*. MACHINE INTELLIGENCE RESEARCH INSTITUTE. https://intelligence.org/files/AIPosNegFactor.pdf

Zhang, Z., Ning, H., Shi, F., Farha, F., Xu, Y., Xu, J., Zhang, F., & Choo, K.-K. R. (2022). Artificial intelligence in cyber security: Research advances, challenges, and opportunities. *Artificial Intelligence Review*, 1-25.

Zheng, H., & Ling, R. (2021). Drivers of social media fatigue: A systematic review. *Telematics and Informatics*, *64*, 101696. https://doi.org/10.1016/j.tele.2021.101696

Zwetsloot, R., & Dafoe, A. (2019, febrero 11). *Thinking About Risks From AI: Accidents, Misuse and Structure*. Lawfare. https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure