analecta polit. | Vol. 14 | No. 26 | PP. 01-25
| enero - junio | 2024 | ISSN-e: 2390-0067 (en línea) | Medellín-Colombia

doi: http://dx.doi.org/10.18566/apolit.v14n26.a08

Revisión sistemática de taxonomías de riesgos asociados a la Inteligencia Artificial

Systematic review of taxonomies of risks associated with Artificial Intelligence

Guillem Bas Graells*
Magíster
Escuela Nacional de Estudios Políticos y Administrativos
Observatorio de Riesgos Catastróficos Globales, USA
ORCID: https://orcid.org/0009-0003-3541-2208
Correo electrónico: gbasg@riesgoscatastroficosglobales.com

Roberto Tinoco Devia*
Magíster
Universidad de los Andes
Observatorio de Riesgos Catastróficos Globales, USA
ORCID: https://orcid.org/0009-0004-7763-6979
Correo electrónico: rtinocod@riesgoscatastroficosglobales.com

Claudette Salinas Leyva*
Abogada
Instituto Tecnológico Autónomo de México (ITAM)
Observatorio de Riesgos Catastróficos Globales, USA
ORCID: https://orcid.org/0009-0009-2625-4563
Correo electrónico: claudette.salinas10@gmail.com

Cómo citar este artículo en APA:
Bas Graells, G.; Tinoco Devia, R.; Salinas Leyva, C. y Sevilla Molina, J. (2024). Revisión sistemática de taxonomías de riesgos asociados a la Inteligencia Artificial. *Analecta Política*, 14(26), 1-25. doi: http://dx.doi.org/10.18566/apolit.v14n26.a08

Fecha de recepción:
26.09.2023
Fecha de aceptación:
07.12.2023

Jaime Sevilla Molina
Matemático
Universidad Complutense de Madrid
Epoch, USA
ORCID: https://orcid.org/my-orcid?orcid=0000-0002-4454-1146
Correo electrónico: jaimesevillamolina@gmail.com

* Estos autores han contribuido de manera equivalente
a este trabajo como autores principales

Resumen

Este artículo realiza una revisión sistemática de treinta y seis taxonomías de riesgos asociados a la Inteligencia Artificial (IA) que se han realizado desde el 2010 hasta la fecha, utilizando como metodología el protocolo *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA). El estudio se basa en la importancia de estas para estructurar la investigación de los riesgos y para distinguir y definir amenazas. Ello permite identificar las cuestiones que generan mayor preocupación y, por lo tanto, requieren mejor gobernanza. La investigación permite extraer tres conclusiones. En primer lugar, se observa que la mayoría de los estudios se centran en amenazas como la privacidad y la desinformación, posiblemente debido a su concreción y evidencia empírica existente. Por el contrario, amenazas como los ciberataques y el desarrollo de tecnologías estratégicas son menos citadas, a pesar de su creciente relevancia. En segundo lugar, encontramos que los artículos enfocados en el origen del riesgo tienden a considerar más frecuentemente riesgos extremos en comparación con los trabajos que abordan las consecuencias. Esto sugiere que la literatura ha sabido identificar las potenciales causas de una catástrofe, pero no las formas concretas en las que esta se puede materializar en la práctica. Finalmente, existe una cierta división entre aquellos artículos que tratan daños tangibles presentes y aquellos que cubren daños potenciales futuros. No obstante, varias amenazas se tratan en la mayoría de los artículos de todo el espectro indicando que existen puntos de unión entre clústeres.

Palabras clave: Inteligencia artificial, Riesgo, Amenaza, Daño, Perjuicio, Taxonomía.

Abstract

This article performs a systematic review of thirty-six taxonomies of risks associated with Artificial Intelligence (AI) that have been conducted from 2010 to date, using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol as a methodology. The study is based on the importance of these to structure risk research and to distinguish and define threats. This makes it possible to identify the issues that are of greatest concern and therefore require better

governance. Three conclusions can be drawn from the research. First, it is observed that most studies focus on threats such as privacy and disinformation, possibly due to their concreteness and existing empirical evidence. In contrast, threats such as cyberattacks and the development of strategic technologies are less cited, despite their increasing relevance. Second, we find that articles focused on the origin of risk tend to consider more frequently extreme risks compared to papers addressing consequences. This suggests that the literature has been able to identify the potential causes of a catastrophe, but not the concrete ways in which it may materialize in practice. Finally, there is some division between those articles that deal with present tangible damage and those that cover potential future damage. Nevertheless, several threats are addressed in the majority of articles across the spectrum indicating that there are commonalities between clusters.

Keywords: Artificial intelligence, Risk, Threat, Damage, Harm, Taxonomy.

Introducción

La inteligencia artificial (IA) promete tener un impacto significativo en la sociedad, tanto positivo como negativo. El potencial transformativo de la tecnología se ha hecho especialmente evidente a raíz de los recientes avances en el campo, que han intensificado la discusión social y académica sobre sus implicaciones. En el apartado de riesgos, muchos han señalado problemas como la amplificación de sesgos, las violaciones de la privacidad o la proliferación de desinformación. En algunos casos, la posibilidad de una IA autónoma, con capacidades avanzadas, se ha considerado, incluso, como una potencial causa de riesgo catastrófico global, equiparable a amenazas como una pandemia o una guerra nuclear (CAIS, 2023).

A través del estudio y comparación de distintas taxonomías, este artículo presenta un análisis de la discusión académica alrededor de los riesgos derivados de la IA. El objetivo principal de este esfuerzo es identificar qué riesgos se plantean más habitualmente y desde qué perspectivas se abordan. Con ello, se pretende reconocer las principales coincidencias y discrepancias en las posiciones de los expertos.

La investigación tiene relevancia tanto teórica como práctica. En primer lugar, compendiar todo el trabajo previo es fundamental para obtener una visión global del estado del arte en la materia. En segundo lugar, la síntesis expuesta en el artículo puede ayudar a identificar los riesgos que se consideran más urgentes y, eventualmente, informar decisiones en el diseño de políticas, regulación y estándares.

Así, el artículo se divide en cinco secciones. En el primer apartado, se exponen trabajos previos que han realizado revisiones de taxonomías de riesgos asociados a la IA. En el segundo, exponemos la metodología empleada en este trabajo para la identificación de artículos, la clasificación de estos, y la extracción de información. En tercer lugar, se presentan los resultados, definiendo las categorías en las que se clasifica el riesgo y la frecuencia con la que aparecen. En cuarto lugar, se discuten las implicaciones de los resultados. Se finaliza el artículo con una conclusión en la que se ofrece un resumen conciso de los resultados más significativos de la investigación, así como sus limitaciones.

Trabajos previos

La taxonomía y análisis de los riesgos asociados con la IA han sido objeto de mucho estudio y discusión en la literatura científica durante los últimos años. No obstante, durante la revisión de la literatura, encontramos solo dos artículos que hacen una revisión sistemática de estas taxonomías, en contraste con realizar una nueva taxonomía independiente:

- The risks associated with Artificial General Intelligence: A systematic review (McLean et al., 2021).
- A Survey of the Potential Long-term Impacts of AI How AI Could Lead to Long-term Changes in Science, Cooperation, Power, Epistemics and Values (Clarke & Whittlestone, 2022).

En el caso de McLean et al. (2021), los autores tuvieron como objetivo resumir y categorizar los riesgos de la inteligencia artificial general (IAG), basándose en la literatura científica revisada por pares. En el proceso de revisión, se identificaron dieciséis artículos elegibles para su inclusión. Los artículos seleccionados abarcan una variedad de enfoques y perspectivas, incluyendo discusiones filosóficas, aplicaciones de técnicas de modelado y evaluaciones de los marcos y procesos actuales en relación con la IAG. Los autores concluyen que las taxonomías analizadas carecen, en general, de técnicas de modelización, y que los riesgos identificados por las taxonomías son generales, pues son muy pocos los artículos que se centran en riesgos de la IAG en dominios más específicos. Asimismo, consideran que no existe consenso en las terminologías utilizadas, lo que lleva, además, a un problema de falta de información en algunos casos.

El otro trabajo relevante para este caso es el de Clarke & Whittlestone (2022). Este realiza una revisión de literatura sobre el potencial impacto de la IA a lar-

go plazo, centrándose en cinco áreas: ciencia, cooperación, poder, epistemología y valores. Según cada dimensión, los autores presentan escenarios hipotéticos sustancialmente mejores y peores que el actual para la humanidad. El artículo concluye que la IA representa una herramienta poderosa que puede moldear el futuro de la sociedad, tanto positiva como negativamente, dependiendo de cómo se gestione su desarrollo y uso. En este sentido, así como se destaca que la IA tiene el potencial de impulsar el progreso en esas cinco categorías, también se advierte sobre los riesgos asociados, tales como asimetrías de la información, conflictos, desigualdad económica, manipulación y cambios en la definición de valores.

Estos dos estudios previos proporcionaron un marco para nuestra investigación, que también se centra en el análisis de los riesgos asociados con la inteligencia artificial a partir de la revisión de literatura secundaria.

A continuación, detallamos la metodología utilizada para este artículo en aras de ampliar el alcance de la revisión de taxonomías e incluir una gama más amplia de literatura, así como para explorar en mayor profundidad los riesgos específicos asociados con la implementación de la IA en diferentes dominios.

Metodología

A partir de McLean et al. (2021) y Clarke & Whittlestone (2022), este artículo busca ampliar el alcance de la revisión para incluir más literatura y explorar en mayor profundidad los riesgos específicos asociados con la implementación de la IA en diferentes dominios. El proceso de investigación en el que se sustenta este artículo se inició con una revisión sistemática de la literatura académica basada en el protocolo *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) (Page et al., 2021). Estas son una serie de pautas reconocidas internacionalmente para mejorar la calidad y la transparencia de las revisiones sistemáticas y los metaanálisis.

La búsqueda se realizó en julio de 2023. Los artículos se extrajeron de seis repositorios electrónicos: IEEE Explore (n = 200), Phil Papers (n = 200), ScienceDirect (n = 200), arXiv (n = 192), ACM Digital Library (n = 106) y Taylor&Francis (n = 45). Los tres primeros arrojaron más resultados, pero, para limitar el alcance del proceso, solo se revisaron los doscientos primeros al ordenarlos por relevancia.

La consulta de búsqueda efectuada se limitó a términos incluidos en el título de la publicación y fue la siguiente:

> ("Artificial intelligence" OR AI OR "artificial general intelligence" OR AGI OR superintelligence OR "foundation model*" OR "language model*") AND (Risk* OR danger* OR threat* OR harm* OR impact* OR catastroph* OR malicious OR challenge*) AND NOT (cyber* OR climat* OR bio* OR health* OR medic* OR financ* OR business* OR education* OR COVID-19 OR cancer OR environment* OR explainab* OR market* OR econom* OR psycholog*).[1]

La búsqueda cuenta con tres elementos. En el primero, se listan términos relacionados con la inteligencia artificial, incluyendo conceptos relacionados con futuros sistemas de IA avanzados –"artificial general intelligence", "AGI", "superintelligence"– y los tipos de sistemas actuales más notorios –"foundation model*", "language model*"–. En el segundo, se incluyen conceptos que connotan consecuencias negativas de estos sistemas. Finalmente, se excluyen artículos enfocados en una sola disciplina, como la medicina o la economía.[2]

Para filtrar los resultados, se tuvieron en cuenta diversos criterios de inclusión. Todas las piezas seleccionadas son artículos académicos en inglés,[3] publicados a partir del 2010, que presentan una lista exhaustiva de riesgos asociados a la inteligencia artificial, independientemente de la naturaleza y escala de estos riesgos. Se excluyeron aquellos artículos exclusivamente enfocados en problemas técnicos o en una sola disciplina,[4] así como aquellos que adoptan íntegramente taxonomías ya existentes o que solamente mencionan los riesgos superficialmente y con detalles insuficientes. En casos en los que uno o varios autores presentaron múltiples versiones de una misma categorización, se seleccionó la versión más completa. También se excluyeron aquellos artículos que listan riesgos para argumentar que

1 En el caso de ACM Digital Library, los términos de exclusión no se incluyeron por falta de precisión en la búsqueda resultante.

2 Se excluyeron artículos limitados a una sola disciplina para priorizar aquellos con una visión amplia del panorama general de riesgos, facilitando la comparación. No obstante, se reconoce que algunos de los artículos excluidos pueden abordar riesgos concretos de su correspondiente disciplina y que, por lo tanto, la presente revisión sistemática podría tender a subestimar el volumen de discusión sobre riesgos que se cubren más habitualmente en trabajos especializados. A modo ilustrativo, en arXiv, los operadores NOT excluyeron dieciséis resultados con términos relacionados con la medicina ("health*", "medic*", "COVID-19", "cancer"); doce con el término "cyber*"; once con el término "explainab*"; y once con términos relacionados con la economía ("econom*", "market*", "financ*").

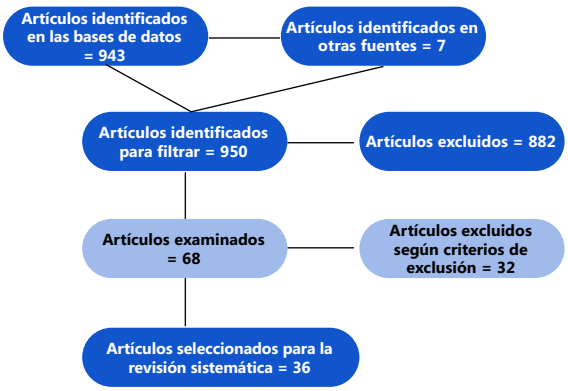
3 Se seleccionaron artículos en inglés por ser más accesibles para los autores y conformar el grueso de la literatura académica sobre el tema. Sin embargo, se reconoce que este criterio puede generar ciertos sesgos y obviar riesgos cubiertos por comunidades no angloparlantes.

4 A pesar de los términos de exclusión en la consulta de búsqueda, los artículos enfocados en una sola disciplina siguieron siendo mayoritarios en esta fase, ya que muchos de ellos contienen términos especializados de la disciplina que no se recogen en la consulta.

estos son generalmente sobreestimados. Esta exclusión se debe al hecho de que su posición es diametralmente opuesta a la de los artículos seleccionados, y no a una negación de la validez del escepticismo.

La figura 1 detalla el proceso de selección de artículos. En la primera fase, novecientos cincuenta artículos fueron filtrados según el título y el resumen del artículo. En la segunda fase, sesenta y ocho artículos fueron filtrados después de una lectura superficial de cada uno. Finalmente, treinta y seis fueron seleccionados para la revisión sistemática.

Figura 1. Proceso de selección de artículos siguiendo el protocolo PRISMA



Fuente: Elaboración propia.

Luego se dividieron los artículos seleccionados en dos grupos según el enfoque adoptado. Las taxonomías por amenaza son aquellas cuyas categorías conciernen al impacto del riesgo. Esto incluye el evento a través del cual se materializa, como un ataque cibernético o una campaña de desinformación; la naturaleza del escenario consecuente, como una mayor desigualdad socioeconómica o una mayor toxicidad en la conversación pública; y el valor afectado por el riesgo, como en el caso de las violaciones de la privacidad. Las taxonomías por vector de origen son aquellas cuyas categorías se refieren a la causa del riesgo. Esto incluye factores relacionados con el funcionamiento de los sistemas de IA, el uso de estos sistemas y otros factores sistémicos.

Una vez hecha esta distinción, se procedió a la extracción de información y evaluación de cada artículo revisado. En ambos casos, la lista de categorías se creó con base en los temas que aparecían en los artículos. Estos fueron actualizándose, incluyendo nuevas categorías o modificando las existentes, a medida que surgían

nuevas cuestiones. Los listados finales de amenazas y vectores de origen son el resultado de este proceso iterativo. Para asegurar la recopilación sistemática de toda la información relevante, se creó una plantilla en la que se especifica qué artículos tratan cada amenaza (tabla 1) o vector de origen (tabla 2).

Tabla 1. Listado de los artículos por amenazas

	Título del artículo
1	Evaluating the Social Impact of Generative AI Systems in Systems and Society
2	AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms
3	Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback
4	Current and Near-Term AI as a Potential Existential Risk Factor
5	Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned
6	A Survey on the Potential Long-Term Impacts of AI
7	On the Opportunities and Risks of Foundation Models
8	The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation
9	AI Risk Mitigation Through Democratic Governance: Introducing the 7-Dimensional AI Risk Horizon
10	Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models
11	On pitfalls (and advantages) of sophisticated large language models
12	Ethics of Artificial Intelligence and Robotics
13	Impact of artificial intelligence on civilization: Future perspectives
14	Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI
15	Ethical and social risks of harm from Language Models
16	Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks
17	Amplifying Limitations, Harms and Risks of Large Language Models
18	Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence
19	On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?
20	Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction
21	Harms of AI
22	Artificial Intelligence: Risks to Privacy and Democracy
23	GPT-4 Technical Report
24	Ethical Risk Factors and Mechanisms in Artificial Intelligence Decision Making
25	Managing the ethical and risk implications of rapid advances in artificial intelligence: A literature review
26	Classification of global catastrophic risks connected with artificial intelligence
27	An Overview of Catastrophic AI Risks

Fuente: Elaboración propia.

Tabla 2. Listado de los artículos por vectores de origen

	Título del artículo
1	TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI
2	An Overview of Catastrophic AI Risks
3	Examining the Differential Risk from High-Level Artificial Intelligence and the Question of Control
4	Taxonomy of Pathways to Dangerous AI
5	The Risks of Low Level Narrow Artificial Intelligence
6	How does Artificial Intelligence Pose an Existential Risk?
7	Heterogeneity of AI-Induced Societal Harms and the Failure of Omnibus AI Laws
8	The risks associated with Artificial General Intelligence: A systematic review
9	The Deficiency of "Redline/Greenline" Approach to Risk Management in AI Applications
10	Harms from increasingly agentic algorithmic systems
11	Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies
12	Classification of global catastrophic risks connected with artificial intelligence

Fuente: Elaboración propia.

Asimismo, para cada caso se incluye evidencia en forma del segmento del artículo donde se cubre la categoría en cuestión. Este proceso fue realizado por los tres autores principales de este artículo para minimizar errores y sesgos en la interpretación de la información. Al final, se cuantificó el número de artículos seleccionados que cubre cada categoría. En el análisis posterior, solamente se incluyeron aquellas categorías con un mínimo de cinco menciones, tanto para las amenazas como para los vectores de origen.

Las definiciones de cada categoría, recogidas en la siguiente sección, han sido redactadas de modo que abarquen los planteamientos de todos los artículos incluidos en dicha categoría. No obstante, reconocemos que la categorización de distintas perspectivas supone una cierta simplificación de estos puntos de vista, por lo que algunos matices se pueden diluir. Como tal, el delineamiento final de las categorías podría haber variado ligeramente de haber interpretado estas sutilezas de forma distinta.

Resultados

En esta sección se presentan los resultados obtenidos en el análisis de las taxonomías por amenaza y por vector de origen.

Por amenaza

En la figura 2, se consideraron veintisiete artículos revisados por amenaza. De esta, se desprenden categorías tales como los ciberataques, la militarización, las tecnologías estratégicas, la manipulación, la desinformación, la vigilancia, la desigualdad, el desempleo, la discriminación, el desempoderamiento, las afectaciones medioambientales, la privacidad y la toxicidad.

Figura 2. Matriz de menciones por amenaza

Estudio	Ciberataques	Militarización	Tec. Est.	Manipulación	Desinform.	Vigilancia	Desigualdad	Desempleo	Discriminación	Desempodera.	Medio amb.	Privacidadd	Toxicidad
Solaiman *et al.* (2023)													
Buçinca *et al.* (2023)													
Kirk *et al.* (2023)													
Bucknall & Dori-Dacohoen (2022)													
Ganguli *et al.* (2022)													
Clarke & Whittlestone													
Bommasani *et al.* (2022)													
Brundage *et al.* (2018)													
Garvey (2018)													
Tamkin *et al.* (2021)													
Strasser (2023)													
Müller (2020)													
Rajendra *et al.* (2022)													
Blauth *et al.* (2022)													
Weidinger *et al.* (2021)													
Vasnic-Alujevic *et al.* (2020)													
O'Neil & Connor (2023)													
Hagerty & Rubinov (2019)													
Bender *et al.* (2021)													
Shelby *et al.* (2023)													
Acemoglu (2021)													
Manheim & Kaplan (2019)													
OpenAI (2023)													
Guan *et al.* (2022)													
Meek *et al.* (2017)													
Turchin & Denkenberger (2018)													
Hendrycks *et al.* (223)													
Total (#)	9	12	8	19	20	14	15	18	18	18	11	20	12
Prevalencia (%)	33%	44%	30%	70%	74%	52%	56%	67%	67%	67%	41%	74%	44%

El estudio sí describe la amenaza
El estudio no describe la amenaza

Fuente: Elaboración propia.

La tabla 3 incluye todas estas categorías, una descripción de cada una y ejemplos extraídos de artículos revisados.

Tabla 3. Clasificación de amenazas

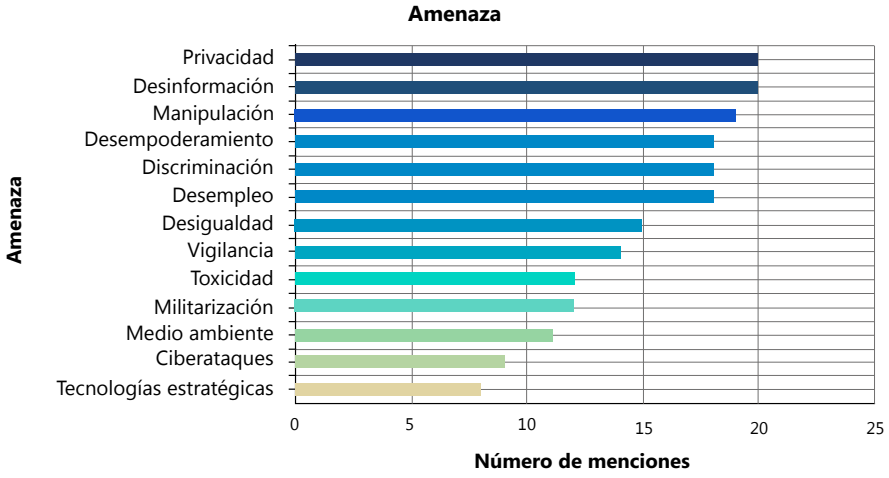
Categoría	Descripción
Ciberataques	Automatización y mejora de tareas relevantes para la ejecución de ataques cibernéticos. Por ejemplo, GPT-4 es útil para algunas subtareas de ingeniería social como redactar correos electrónicos de *phishing*, y para explicar vulnerabilidades de *software* (OpenAI, 2023).
Militarización	Utilización de armas autónomas letales diseñadas para seleccionar y atacar objetivos sin la necesidad de control humano, lo cual puede facilitar masacres y otros daños a gran escala (Blauth et al., 2022). Automatización de los procesos de toma de decisión en el ámbito militar, que puede facilitar que los conflictos escalen más rápidamente (Clarke & Whittlestone, 2022; Hendrycks et al., 2023).
Tecnologías estratégicas	Utilización del desarrollo científico-tecnológico para fines peligrosos o éticamente cuestionables. El diseño de armas biológicas y químicas es el ejemplo más citado (Bucknall & Dori-Hacohen, 2022; Clarke & Whittlestone, 2022; Hendrycks et al., 2023; OpenAI, 2023).
Manipulación	Influencia en el comportamiento humano a través de la persuasión. Por ejemplo, hacer que las personas sean más propensas a comprar determinados productos o incluso fomentar la autolesión (Bommasani et al., 2022; Brundage et al., 2018; Müller, 2020; Rajendra et al., 2022; Solaiman et al., 2023).
Desinformación	Generación de información falsa o engañosa a gran escala. Por ejemplo, los votantes indecisos pueden ser dirigidos con mensajes personalizados para afectar su comportamiento electoral (Brundage et al., 2018).
Vigilancia	Utilización de herramientas de IA por parte de gobiernos y empresas para llevar a cabo vigilancia masiva. Por ejemplo, los sistemas de reconocimiento facial pueden servir como herramienta de control social, mientras que los modelos de lenguaje se pueden utilizar para vigilar la comunicación de texto en entornos laborales, sociales y otros (Acemoglu, 2021; Solaiman et al., 2023; Vesnic-Alujevic et al., 2020).
Desigualdad	Acceso dispar a los recursos necesarios para el desarrollo y uso de la inteligencia artificial, agravando la marginación de las comunidades desfavorecidas (Clarke & Whittlestone, 2022; Solaiman et al., 2023) e incrementando el poder de los desarrolladores y proveedores de la tecnología (Bucknall & Dori-Hacohen, 2022; Hendrycks et al., 2023).

Categoría	Descripción
Desempleo	Automatización masiva de tareas, causando el reemplazo de trabajadores humanos en todo el mercado laboral. Las corporaciones enfrentarán incentivos significativos para automatizar el trabajo humano, lo que podría conducir a un desempleo masivo (Hendrycks et al., 2023).
Discriminación por sesgo	Perpetuación e intensificación de sesgos existentes en la sociedad si los datos de entrenamiento o el proceso de desarrollo reflejan tales desigualdades. Por ejemplo, la clasificación de imágenes por IA que se produce en el *software* de reconocimiento facial genera resultados discriminatorios (Manheim & Lyric Kaplan, 2018), mientras que la IA generativa puede sobrerrepresentar, subrepresentar o estereotipar a determinados grupos sociales (Bommasani et al., 2022).
Desempoderamiento	Pérdida de la agencia humana causada por el abandono progresivo del poder de toma de decisión y la consecuente dependencia de la tecnología. A medida que los usuarios se sientan más cómodos con el sistema, la dependencia del modelo puede obstaculizar el desarrollo de nuevas habilidades o incluso conducir a la pérdida de habilidades importantes. A medida que crece la confianza general en el modelo, es menos probable que los usuarios cuestionen o verifiquen las respuestas del modelo (Acemoglu, 2021; O'Neill & Connor, 2023; OpenAI, 2023).
Afectaciones medioambien-tales	Generación de residuos electrónicos y uso de gran cantidad de energía, con externalidades para el medioambiente. Según algunos autores, estos costos ambientales castigan sobre todo a las comunidades marginadas (Bender et al., 2021).
Privacidad	Recopilación, almacenamiento y análisis no autorizados de grandes cantidades de datos personales. Algunos modelos, por ejemplo, cuentan en sus bases de datos con documentos altamente confidenciales o información de identificación personal, como números de teléfono, direcciones y registros médicos privados (Kirk et al., 2023; Müller, 2020; OpenAI, 2023; Solaiman et al., 2023; Strasser, 2023).
Toxicidad	Propagación de contenido ofensivo. Los modelos de lenguaje, por ejemplo, pueden generar mensajes que representen abuso sexual, violencia, racismo o sexismo (Strasser, 2023).

Fuente: Elaboración propia.

De dichas categorías, las amenazas de desinformación y privacidad son las más frecuentes, cada una con veinte menciones (74 %). Así, la figura 3 sugiere una alta preocupación en la literatura revisada sobre cómo la IA puede socavar la confianza y la integridad de los datos personales. Por otro lado, destaca que las amenazas concernientes a la seguridad son las menos mencionadas. Este es el caso de la militarización (44 %), los ciberataques (33 %) y el desarrollo de tecnologías estratégicas (30 %).

Figura 3. Lista de amenazas por número de menciones



Fuente: Elaboración propia.

Entre las amenazas que no entraron en el análisis por número insuficiente de menciones, destacan las violaciones de la propiedad intelectual (Bommasani et al., 2022; O'Neill & Connor, 2023; Solaiman et al., 2023; Strasser, 2023); la erosión y polarización política (Acemoglu, 2021; Manheim & Lyric Kaplan, 2018; Meek et al., 2016; Shelby et al., 2023); la falta de responsabilidad legal de las decisiones automatizadas (Buçinca et al., 2023; Guan et al., 2022; Meek et al., 2016; Vesnic-Alujevic et al., 2020); y un aceleramiento repentino de procesos, inasumible para las estructuras sociales actuales (Blauth et al., 2022; Clarke & Whittlestone, 2022).

Por vector de origen

En la figura 4 se seleccionaron once artículos que clasifican los riesgos según su vector de origen. Estos incluyen el desalineamiento, el uso malicioso, los riesgos estructurales, los riesgos accidentales y las dinámicas competitivas.

Figura 4. Matriz de menciones por vectores de origen

Estudio	Desalineamiento	Uso malicioso	Riesgos estructurales	Riesgos accidentales	Dinámicas competitivas
Critch, A. & Russell, S. (2023)					
Hendrycks et al. (2022)					
Kilian et al. (2022)					
Yampolskiy (2015)					
Page et al. (2018)					
Vold Y Harris (2021)					
McLean et al. (2023)					
Kuleshov et al. (2021)					
Chan et al. (2023)					
Scherer (2015)					
Turchin & Denkenberger (2018)					
Total (#)	10	9	5	6	7
Prevalencia (%)	83%	75%	42%	50%	58%

El estudio sí describe el vector de origen	
El estudio no describe el vector de origen	

Fuente: Elaboración propia.

El factor común de estas taxonomías es que se enfocan en la causa del riesgo, en lugar de su consecuencia. De los artículos analizados, en la Tabla 4 se pueden distinguir las categorías y sus respectivas definiciones.

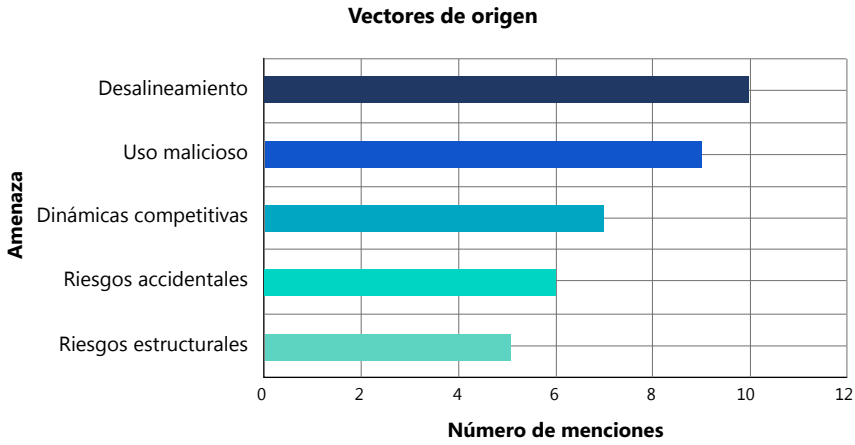
Tabla 4. Clasificación de riesgos según su vector de origen

Categoría	Descripción
Desalineamiento	Situaciones en las que los sistemas de IA actúan competentemente, pero de un modo distinto al que pretendían sus desarrolladores. La generalización errónea de objetivos y el hackeo de recompensas son las manifestaciones más mencionadas de desalineamiento (Kilian et al., 2023; Scherer, 2015; Turchin & Denkenberger, 2020; Vold & Harris, 2021; Yampolskiy, 2015). En algunos casos, el desalineamiento se vincula con la posibilidad de que un sistema adquiera autónomamente recursos útiles para perseguir sus objetivos y, por lo tanto, sea más difícil de controlar (Hendrycks et al., 2023).
Dinámicas competitivas	Presiones que incentivan el aceleramiento del desarrollo y/o despliegue de la IA en detrimento de medidas de precaución. La posibilidad de una carrera por la IA entre empresas, gobiernos y ejércitos es el ejemplo más citado (Chan et al., 2023; Critch & Russell, 2023; Hendrycks et al., 2023; Turchin & Denkenberger, 2020; Vold & Harris, 2021).
Riesgos accidentales	Procesos automatizados defectuosos que no necesariamente implican desalineamiento, sino un rendimiento subóptimo con causas diversas: errores en la interpretación del entorno, vulnerabilidades ante ataques adversarios, excesiva complejidad, etc. Un ejemplo extremo de un riesgo accidental sería un error de percepción en un radar de misiles nucleares, que podría causar una catástrofe si la detección de un ataque enemigo desencadena automáticamente la retaliación (Turchin & Denkenberger, 2020). También pueden ocurrir accidentes organizacionales que impliquen filtraciones al público o robos por actores malintencionados (Hendrycks et al., 2023).
Riesgos estructurales	Impacto sistémico del despliegue a gran escala de la IA, surgido, entre otros, de la interacción disruptiva de la tecnología con el entorno (Chan et al., 2023; Kilian et al., 2023; Vold & Harris, 2021) o de situaciones de tragedia de los comunes (Critch & Russell, 2023; Turchin & Denkenberger, 2020).
Uso malicioso	Aprovechamiento de la IA por parte de actores malintencionados, estatales y no estatales, para causar daño. Los ciberataques automatizados, *deep fakes* y el uso de armas autónomas letales son algunos ejemplos (Hendrycks et al., 2023; Kilian et al., 2023; Page et al., 2018; Vold & Harris, 2021).

Fuente: Elaboración propia.

De las categorías mencionadas, en la Figura 5 se puede observar que el desalineamiento y el uso malicioso son las dos categorías más discutidas, con 10 y 9 menciones, respectivamente. Las dinámicas competitivas, los riesgos accidentales y los riesgos estructurales quedaron por detrás, con 7, 6 y 5 menciones cada uno.

Figura 5. Lista de vectores de origen por número de menciones



Fuente: Elaboración propia.

Algunos vectores de origen y factores agravantes del riesgo que no entraron en el análisis por número insuficiente de menciones son la opacidad (Kuleshov et al., 2021; Scherer, 2015); la falta de robustez ante ataques adversarios y cambios en el entorno (Page et al., 2018); y el manejo inadecuado de la IA (McLean et al., 2021; Scherer, 2015).

Discusión

La IA es un campo vasto y diverso, y las taxonomías de riesgos proporcionan una estructura que puede guiar la investigación en áreas específicas. El análisis comparativo de taxonomías de riesgos ayuda a distinguir los diferentes tipos de amenazas y causas, creando marcos útiles para la comprensión de los desafíos derivados de la tecnología. Eventualmente, un marco bien estructurado y definido

puede fundamentar el desarrollo de estándares, regulaciones y otras medidas de gobernanza.

La revisión sistemática realizada en este artículo arroja claridad sobre las distintas formas en las que se investigan los riesgos asociados a la IA. Por un lado, la mayoría de los artículos por amenazas se enfocan en riesgos relativamente menos graves, como la privacidad o la desinformación. Atribuimos esta tendencia a dos factores. En primer lugar, este tipo de amenazas es más fácil de identificar y concretar, por lo que se perciben como más probables o menos especulativas que otras. En algunos casos, como el de la privacidad o la discriminación, ya existe evidencia del daño causado. No obstante, sería beneficioso que las evaluaciones del riesgo estuvieran basadas en un ejercicio de prospección que ayudara a identificar riesgos futuros, algunos de los cuales ya están basados en indicios empíricos. En segundo lugar, muchos de los artículos revisados se limitan a los modelos de lenguaje,[5] por lo que los riesgos asociados con la información pueden aparecer sobrerrepresentados. Resulta importante que, a pesar de que los sistemas de IA más avanzados de la actualidad sean modelos de lenguaje, las taxonomías del riesgo consideren también los riesgos derivados de otros tipos de sistemas.

Cabe destacar que la intensificación de los ciberataques y el desarrollo de tecnologías estratégicas son las amenazas menos citadas en los artículos revisados. Esto contrasta con la creciente evidencia de que los modelos actuales pueden ayudar a diseñar patógenos y toxinas (Soice et al., 2023; Urbina et al., 2022) o ejecutar ataques cibernéticos (Guembe et al., 2022; Hazell, 2023). Resulta importante que los modelos de amenazas se actualicen para integrar el riesgo asociado al incremento de las capacidades de los sistemas de IA.

Por otro lado, la mayoría de los artículos por vector de origen consideran riesgos más extremos, mayoritariamente imputados al desalineamiento y el uso malicioso. En concreto, un 73 % de estos artículos hace referencia a "riesgos catastróficos" o "riesgos existenciales" en el título o el resumen, una cifra que contrasta con el 15 % para el caso de los artículos por amenaza.

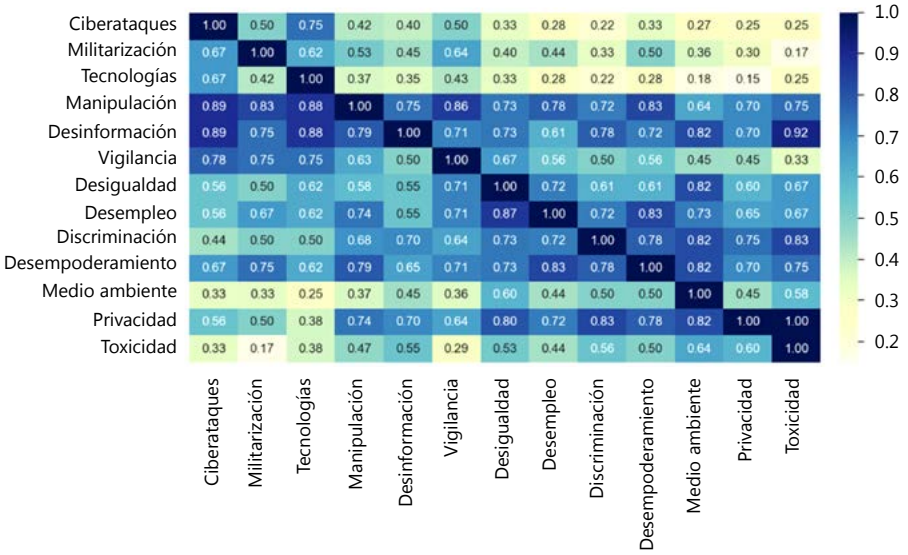
Esta realidad parece sugerir que la literatura académica ha sabido identificar las potenciales causas de una catástrofe, pero no las formas concretas en las que esta se puede materializar en la práctica. Solamente dos artículos que consideran los riesgos catastróficos son exhaustivos en la clasificación tanto de amenazas

5 Ocho artículos se enfocan exclusivamente en los modelos de lenguaje, mientras que otros dos se limitan a la IA generativa y a modelos fundacionales.

como de vectores de origen (Hendrycks et al., 2023; Turchin & Denkenberger, 2020). Si bien las amenazas representadas por futuros sistemas de IA son difíciles de predecir, la literatura se beneficiaría de mayores esfuerzos para identificar y definir posibles daños. El diseño de metodologías para ello podría ser un objeto de estudio relevante para futuros trabajos.

En cuanto a la interacción entre artículos, la figura 6 provee detalles sobre la coocurrencia de distintas amenazas. Las amenazas listadas en el eje vertical se cubren en el correspondiente porcentaje de artículos que también cubren la respectiva amenaza del eje horizontal. Por ejemplo, los ciberataques aparecen mencionados en el 50 % de artículos que cubren militarización, el 75 % de los que cubren tecnologías estratégicas, el 42 % de los que abordan manipulación, etc. Por el otro lado, los artículos que plantean las amenazas del eje horizontal también mencionan las del eje vertical con la frecuencia correspondiente. Por ejemplo, los artículos que cubren ciberataques hacen lo propio con militarización (un 67 % de las veces), tecnologías estratégicas (67 %), manipulación (89 %), etc.

Figura 6. Matriz de coocurrencia entre amenazas y vectores de origen



Fuente: Elaboración propia.

Como se muestra en la matriz, existe cierta división entre aquellos que tratan daños tangibles presentes y aquellos que cubren daños potenciales futuros. En

concreto, las amenazas que más claramente conciernen la seguridad –militarización, ciberataques, tecnologías estratégicas– tienden a coocurrir más frecuentemente, mientras que aquellas asociadas a cuestiones éticas –privacidad, discriminación, toxicidad, etc.– también coinciden más entre ellas.

No obstante, algunas amenazas generan un amplio consenso a lo largo de todo el espectro. El riesgo de manipulación ocurre entre un 64 % y un 89 % con todo el resto de las amenazas. Simultáneamente, el riesgo de desinformación lo hace entre un 61 % y un 92 %, y el riesgo de desempoderamiento entre un 62 % y un 83 %. La existencia de estas categorías conjuntas sugiere que existen puntos de unión entre ambos clústeres. En particular, las tres amenazas presentan connotaciones psicológicas, ya que implican persuasión, engaño y enajenación.

Conclusión

La IA ha emergido como una herramienta transformadora en múltiples dominios, ofreciendo capacidades sin precedentes, pero también presentando riesgos significativos. A través de nuestra revisión sistemática, hemos identificado una serie de riesgos por amenazas y por vectores de origen que reflejan las preocupaciones actuales en la literatura académica sobre la IA.

Debido a la limitación del idioma escogido, una futura línea de investigación incluye la exploración de los riesgos y los vectores de origen que han sido identificados en repositorios y artículos en español, en relación con sus contrapartes angloparlantes. Este enfoque permitiría un análisis académico detallado de las similitudes y diferencias en las taxonomías según los diferentes contextos lingüísticos y culturales.

Es evidente que ciertas amenazas, como la desinformación y la privacidad, son ampliamente reconocidas. Al contrario, otras como la militarización y los ciberataques no reciben la atención equivalente a su potencial impacto. Esta discrepancia puede ser atribuida a la facilidad con la que ciertos riesgos pueden ser identificados y discutidos en comparación con otros.

Además, la literatura parece estar dividida entre abordar daños tangibles actuales y daños potenciales futuros. Esta división sugiere una necesidad de un enfoque más holístico que pueda abordar tanto los riesgos inmediatos como aquellos a largo plazo de la IA. Es esencial que la comunidad académica y la industria

trabajen juntas para anticipar y mitigar dichos riesgos, especialmente a medida que la IA continúa evolucionando y encontrando aplicaciones en áreas críticas.

Las amenazas que coocurren con frecuencia, como la manipulación, la desinformación y el desempoderamiento, indican áreas donde la IA tiene un impacto más profundo en los individuos. Estas amenazas, que involucran persuasión, engaño y enajenación, subrayan la necesidad de abordar no solo los riesgos técnicos de la IA, sino también sus implicaciones éticas y sociales.

En conclusión, la literatura actual proporciona una base sólida de las diferentes taxonomías de riesgos. Sin embargo, es esencial que se continúe expandiendo y refinando la comprensión de estos riesgos para garantizar un marco sobre el cual se puedan desarrollar estándares legales, sociales, éticos y empresariales, asegurando que se cubran los aspectos más relevantes del campo.

Agradecimientos

Nos gustaría agradecer a Pablo Villalobos, Carlos Ignacio Gutiérrez y José Hernández Orallo por su útil discusión y comentarios sobre varias versiones del artículo. Todos los errores restantes son responsabilidad de los autores.

Referencias

Acemoglu, D. (2021). *Harms of AI* (Working Paper 29247). National Bureau of Economic Research. https://doi.org/10.3386/w29247

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623. https://doi.org/10.1145/3442188.3445922

Blauth, T. F., Gstrein, O. J., & Zwitter, A. (2022). Artificial intelligence crime: An overview of malicious use and abuse of AI. *IEEE Access, 10*, 77110-77122. https://ieeexplore.ieee.org/document/9831441

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. https://doi.org/10.48550/arXiv.2108.07258

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*. https://doi.org/10.48550/arXiv.1802.07228

Buçinca, Z., Pham, C. M., Jakesch, M., Ribeiro, M. T., Olteanu, A., & Amershi, S. (2023). AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. *arXiv preprint arXiv:2306.03280*. https://doi.org/10.48550/arXiv.2306.03280

Bucknall, B. S., & Dori-Hacohen, S. (2022). Current and near-term AI as a potential existential risk factor. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 119-129). https://doi.org/10.1145/3514094.3534146

CAIS (2023). *Statement on AI Risk*. Center of AI Safety. https://www.safe.ai/statement-on-ai-risk

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., … Maharaj, T. (2023). Harms from Increasingly Agentic Algorithmic Systems. In *2023 ACM Conference on Fairness, Accountability, and Transparency* (pp.0 651-666). https://doi.org/10.1145/3593013.3594033

Clarke, S. & Whittlestone, J. (2022). A Survey of the Potential Long-term Impacts of AI. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 192-202). https://doi.org/10.1145/3514094.3534131

Critch, A., & Russell, S. (2023). TASRA: A Taxonomy and Analysis of Societal-Scale Risks from AI. *arXiv preprint arXiv:2306.06924*. https://doi.org/10.48550/arXiv.2306.06924

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., ... & Clark, J. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*. https://n9.cl/u57rh

Garvey, C. (2018). AI Risk Mitigation Through Democratic Governance: Introducing the 7-Dimensional AI Risk Horizon. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 366-367). https://doi.org/10.1145/3278721.3278801

Guan, H., Dong, L., & Zhao, A. (2022). Ethical Risk Factors and Mechanisms in Artificial Intelligence Decision Making. *Behavioral Sciences*, *12*(9), Article 9. https://doi.org/10.3390/bs12090343

Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). The Emerging Threat of Ai-driven Cyber Attacks: A Review. *Applied Artificial Intelligence*, *36*(1), 2037254. https://doi.org/10.1080/08839514.2022.2037254

Hagerty, A. & Rubinov, I. (2019). Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. *arXiv preprint arXiv:1907.07892*. https://doi.org/10.48550/arXiv.1907.07892

Hazell, J. (2023). Large language models can be used to effectively scale spear phishing campaigns. *arXiv preprint arXiv:2305.06972*. https://n9.cl/siuce

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An Overview of Catastrophic AI Risks. *arXiv preprint arXiv:2306.12001*. https://doi.org/10.48550/arXiv.2306.12001

Kilian, K. A., Ventura, C. J., & Bailey, M. M. (2023). Examining the Differential Risk from High-level Artificial Intelligence and the Question of Control. *Futures*, *151*, 103182. https://doi.org/10.1016/j.futures.2023.103182

Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2023). Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*. https://doi.org/10.48550/arXiv.2303.05453

Kuleshov, A., Ignatiev, A., & Abramova, A. (2021). The Deficiency of "Redline/Greenline" Approach to Risk Management in AI Applications. In *2021 International Conference Engineering Technologies and Computer Science (EnT)* (pp. 49-55). https://doi.org/10.1109/EnT52731.2021.00015

Manheim, Karl M. and Kaplan, Lyric, Artificial Intelligence: Risks to Privacy and Democracy (October 25, 2018). 21 Yale Journal of Law and Technology 106 (2019), Loyola

Law School, Los Angeles Legal Studies Research Paper No. 2018-37, Available at SSRN: https://ssrn.com/abstract=3273016

McLean, S., Gemma J. M. Read, Jason Thompson, Chris Baber, Neville A. Stanton & Paul M. Salmon (2023) The risks associated with Artificial General Intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence, 35*(5), 649-663. https://doi.org/10.1080/0952813X.2021.1964003

Meek, T., Barham, H., Beltaif, N., Kaadoor, A., & Akhter, T. (2016). Managing the ethical and risk implications of rapid advances in artificial intelligence: A literature review. In *2016 Portland International Conference on Management of Engineering and Technology (PICMET)* (pp. 682-693). https://doi.org/10.1109/PICMET.2016.7806752

Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. En E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (pp. 1-70). CSLI, Stanford University.

O'Neill, M. & Connor, M. (2023). Amplifying Limitations, Harms and Risks of Large Language Models. *arXiv. (arXiv:2307.04821)*. https://doi.org/10.48550/arXiv.2307.04821

OpenAI (2023, marzo 15). *GPT-4 Technical Report*. arXiv.Org. https://n9.cl/zfb8z

Page, J., Bain, M., & Mukhlish, F. (2018, August). The risks of low level narrow artificial intelligence. In *2018 IEEE international conference on intelligence and safety for robotics (ISR)* (pp. 1-6). https://ieeexplore.ieee.org/document/8535903/

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1), 89. https://n9.cl/hgnin

Rajendra, P., Kumari, M., Rani, S., Dogra, N., Boadh, R., Kumar, A., & Dahiya, M. (2022). Impact of artificial intelligence on civilization: Future perspectives. *Materials Today: Proceedings*, *56*, 252-256. https://doi.org/10.1016/j.matpr.2022.01.113

Scherer, M. U. (2015). *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies* (SSRN Scholarly Paper 2609777). https://n9.cl/muwp6

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., … & Virk, G. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 723-741). https://doi.org/10.48550/arXiv.2210.05791

Soice, E. H., Rocha, R., Cordova, K., Specter, M., & Esvelt, K. M. (2023). Can large language models democratize access to dual-use biotechnology? *arXiv preprint arXiv:2306.03809*. https://arxiv.org/abs/2306.03809v1

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., … & Vassilev, A. (2023). Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv preprint arXiv:2306.05949*. https://arxiv.org/abs/2306.05949v2

Strasser, A. (2023). On pitfalls (and advantages) of sophisticated large language models. *arXiv preprint arXiv:2303.17511*. https://doi.org/10.48550/arXiv.2303.17511

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*. https://doi.org/10.48550/arXiv.2102.02503

Turchin, A. & Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *AI & Society*, *35*(1), 147-163. https://n9.cl/uob2z

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

Vesnic-Alujevic, L., Nascimento, S., & Pólvora, A. (2020). Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks. *Telecommunications Policy*, *44*(6), 101961. https://doi.org/10.1016/j.telpol.2020.101961

Vold, K., & Harris, D. R. (2021). How Does Artificial Intelligence Pose an Existential Risk? En C. Véliz (Ed.), *Oxford Handbook of Digital Ethics*. Oxford University Press.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*. https://doi.org/10.48550/arXiv.2112.04359

Yampolskiy, R. V. (2015). Taxonomy of Pathways to Dangerous Artificial Intelligence. *(arXiv:1511.03246). arXiv.* https://doi.org/10.48550/arXiv.1511.03246

Material suplementario**. Matriz de clasificación**

Lista de artículos evaluados

1. Evaluating the Social Impact of Generative AI Systems in Systems and Society (Solaiman et al., 2023)**.**
2. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms (Buçinca et al., 2023)**.**
3. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback (Kirk et al., 2023)**.**
4. Current and Near-Term AI as a Potential Existential Risk Factor (Bucknall & Dori-Hacohen, 2022).
5. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned (Ganguli et al., 2022)**.**
6. A Survey on the Potential Long-Term Impacts of AI (Clarke & Whittlestone, 2022)**.**
7. On the Opportunities and Risks of Foundation Models (Bommasani et al., 2022)**.** The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation (Brundage et al., 2018)**.**
8. AI Risk Mitigation Through Democratic Governance: Introducing the 7-Dimensional AI Risk Horizon (Garvey, 2018)**.**
9. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models (Tamkin et al., 2021)**.**
10. On pitfalls (and advantages) of sophisticated large language models (Strasser, 2023)**.**
11. Ethics of Artificial Intelligence and Robotics (Müller, 2020)**.**
12. Impact of artificial intelligence on civilization: Future perspectives (Rajendra et al., 2022)**.**
13. Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI (Blauth et al., 2022)**.**
14. Ethical and social risks of harm from Language Models (Weidinger et al., 2021)**.**
15. Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks (Vesnic-Alujevic et al., 2020).
16. Amplifying Limitations, Harms and Risks of Large Language Models (O'Neill & Connor, 2023)**.**

17. Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence (Hagerty & Rubinov, 2019).
18. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? (Bender et al., 2021).
19. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction (Shelby et al., 2023).
20. Harms of AI (Acemoglu, 2021).
21. Artificial Intelligence: Risks to Privacy and Democracy (Manheim & Lyric Kaplan, 2018).
22. GPT-4 Technical Report (OpenAI, 2023).
23. Ethical Risk Factors and Mechanisms in Artificial Intelligence Decision Making (Guan et al., 2022).
24. Managing the ethical and risk implications of rapid advances in artificial intelligence: A literature review (Meek et al., 2016).
25. Classification of global catastrophic risks connected with artificial intelligence (Turchin & Denkenberger, 2020).
26. An Overview of Catastrophic AI Risks (Hendrycks et al., 2023).
27. TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI (Critch & Russell, 2023).
28. Examining the Differential Risk from High-Level Artificial Intelligence and the Question of Control (Kilian et al., 2023)
29. Taxonomy of Pathways to Dangerous AI (Yampolskiy, 2015).
30. The Risks of Low Level Narrow Artificial Intelligence (Page et al., 2018).
31. How does Artificial Intelligence Pose an Existential Risk? (Vold & Harris, 2021).
32. The risks associated with Artificial General Intelligence: A systematic review (McLean et al., 2021).
33. The Deficiency of "Redline/Greenline" Approach to Risk Management in AI Applications (Kuleshov et al., 2021).
34. Harms from increasingly agentic algorithmic systems (Chan et al., 2023).
35. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies (Scherer, 2015).
36. Classification of global catastrophic risks connected with artificial intelligence (Turchin & Denkenberger, 2020).