Question
Periodismo / Comunicación
ISSN 1669-6581

Desafíos de la Inteligencia Artificial generativa. *Tres escalas y dos enfoques transversales*
Flavia Costa, Julián Andrés Mónaco, Alejandro Covello, Iago Novidelsky, Ximena Zabala, Pablo Rodríguez
Question/Cuestión, Nro.76, Vol.3, Diciembre 2023
ISSN: 1669-6581
URL de la Revista: https://perio.unlp.edu.ar/ojs/index.php/question/
IICom -FPyCS -UNLP
DOI: https://doi.org/10.24215/16696581e844

Desafíos de la Inteligencia Artificial generativa
Tres escalas y dos enfoques transversales

Challenges of Generative Artificial Intelligence
Three scales and two transversal approaches

Flavia Costa
Tecnocenolab (UBA) y Conicet
Argentina
flavc@hotmail.com
https://orcid.org/0000-0001-8519-5860

Julián Andrés Mónaco
CONICET-IDAES (UNSAM) / UBA
Argentina
julmonaco@gmail.com
https://orcid.org/0000-0002-1918-2591

Alejandro Covello
Tecnocenolab (UBA) y JST
Argentina
alejandrocovello8@gmail.com

https://orcid.org/0009-0008-2410-939

Iago Novidelsky
Tecnocenolab (UBA) y JST
Argentina
iagonovidelsky@gmail.com
https://orcid.org/0009-0009-6154-809X

Ximena Zabala
Tecnocenolab (UBA)
Argentina
ximenavzabala@gmail.com
https://orcid.org/0009-0008-1302-9159

Pablo Rodríguez
Tecnocenolab (UBA) y Conicet
Argentina
pablomanolorodriguez@gmail.com
https://orcid.org/0000-0003-0605-1899

Resumen

El objetivo de este artículo es ofrecer una perspectiva analítica para comprender y situar en su dimensión específica los diferentes debates que atraviesan la conversación pública sobre las Inteligencias Artificiales generativas y los modelos de lenguaje grandes (LLM, por sus siglas en inglés). En primer lugar, identificamos cinco rasgos de las IA: no son una tecnología, sino que se trata de metatecnologías; constituyen no un dispositivo técnico, sino un mundoambiente; pueden ser tecnologías de alto riesgo y requieren de un tratamiento acorde en su ciclo de vida; las IA generativas y en particular los LLM no son sólo Inteligencia Artificial, sino también Sociedad Artificial; la perspectiva de la ética de la IA no es suficiente para abordarlas y es preciso promover un enfoque desde la ética organizacional de la IA y desde el pensamiento

sistémico. En segundo lugar, recortamos las distintas escalas en las que actualmente se desarrollan las IA: la escala micro, la escala meso (la más propia para situar las políticas públicas) y la escala macro. En tercer lugar, presentamos dos enfoques transversales para el abordaje de las IA: el jurídico, orientado a la responsabilidad, y el sistémico, orientado a la protección y a la seguridad.

Abstract

The objective of this article is to offer an analytical perspective to understand and situate in their specific dimension the different debates that cross the public conversation about generative Artificial Intelligences and large language models (LLM). First, we identify five traits of AI: they are not a technology, but rather meta-technologies; They constitute not a technical device, but a world-environment; They can be high-risk technologies and require appropriate treatment in their life cycle; Generative AI and in particular LLM are not only Artificial Intelligence, but also Artificial Society; The perspective of AI ethics is not sufficient to address them and it is necessary to promote an approach from the organizational ethics of AI and from systemic thinking. Secondly, we cut out the different scales at which AI is currently developed: the micro scale, the meso scale (the most suitable for situating public policies) and the macro scale. Thirdly, we present two transversal approaches to addressing AI: the legal one, oriented towards responsibility, and the systemic one, oriented towards protection and security.

Palabras clave: Inteligencia Artificial; Inteligencia Artificial generativa; riesgos y seguridad de la IA; sociedad artificial

Key words: Artificial intelligence; Generative Artificial Intelligence; AI risks and safety; artificial society

Introducción

Desde finales de 2022, los debates en torno a los posibles efectos de la Inteligencia Artificial (IA) y, en particular, de las IA generativas (IAG) se han situado en el centro de la conversación pública, tanto en el ámbito internacional como en los ámbitos regional y local. En las discusiones respecto de los usos de estos sistemas técnicos intervienen diferentes actores, entre los que se cuentan expertos gubernamentales, científicos, organismos internacionales, representantes de corporaciones de tecnología e información y diferentes actores de la sociedad civil.

Las inteligencias artificiales comenzaron a ser diseñadas y construidas a mediados de los años cincuenta del siglo XX, en un campo que cruza las ciencias de la computación, las ciencias cognitivas, la cibernética y las ciencias de la comunicación, y que investiga y desarrolla tecnologías que replican o emulan ciertos comportamientos humanos. De este modo, hablamos de IA, en principio, cuando un sistema computacional tiene la capacidad de realizar "funciones cognitivas similares a las de los humanos" (OCDE, 2019). (2)

Más cerca en el tiempo, las IA generativas han venido incrementando su potencia de crear, inventar y, sobre todo, operar en el mundo. Su expansión y puesta en disponibilidad masiva, en particular a partir de 2022 con el Chat GPT, suscitó declaraciones de científicos, empresarios y expertos en políticas públicas de diversos países e instituciones internacionales con el propósito de regular sus alcances y alertar sobre sus peligros. Destacan en este sentido los informes de la OCDE (2019b) y de la Unesco (2022), así como los marcos normativos establecidos por los Estados Unidos y la Unión Europea (UE), que proponen dos enfoques muy distintos para abordar la cuestión. (3)

En la Argentina, el desarrollo de la IA también está siendo tema de estudio y análisis, y en los últimos años ha sido puesto en consideración desde el ámbito gubernamental. Existen diferentes proyecciones acerca del impacto económico positivo que puede tener para el país y para la región la incorporación de IA en diferentes procesos productivos: un estudio de 2020 del Banco Interamericano de Desarrollo (BID) proyectaba que esta adopción podía representar para América Latina la oportunidad de elevar en un 14% el Producto Bruto Interno (Gómez Mont *et al.*, 2020). Y un informe del Centro de Implementación de Políticas Públicas para la Equidad y el Crecimiento (CIPPEC) de 2018 señalaba que la adopción de IA podría duplicar la tasa de crecimiento inercial de la economía argentina (Albrieu *et al.*, 2018).

También se desarrollaron recomendaciones no vinculantes para un uso fiable de estas tecnologías: el 2 de junio de 2023 se publicaron en el Boletín Oficial de la Nación las Rec*omendaciones para una inteligencia artificial fiable,* en las que se señala que "la irrupción de la Inteligencia Artificial [...] empuja a los Estados a definir estrategias para encauzar el potencial transformador de esta tecnología en la resolución de problemas concretos y a favor del bien común" (Subsecretaría de Tecnologías de la Información, 2023, 2). El documento, emitido por la Subsecretaría de Tecnologías de la Información, dependiente de la Secretaría de Innovación Pública de la Jefatura de Gabinete de Ministros, establece asimismo una serie de valores de alineamiento, a saber: proporcionalidad e inocuidad; seguridad y protección; equidad y no discriminación; sostenibilidad; derecho a la intimidad y protección de datos; supervisión y decisión humanas; transparencia y explicabilidad; responsabilidad y rendición de cuentas; sensibilización y educación; y gobernanza y colaboración adaptativa y de múltiples partes interesadas. (4)

Esta misma dependencia –que entre sus competencias posee la de "Entender en la ciberseguridad y protección de infraestructuras críticas de información y comunicaciones asociadas del Sector Público Nacional y de los servicios de información y comunicaciones definidos en el artículo 1° de la Ley N° 27.078"– emitió en julio del mismo año una Guía de Notificación y Gestión de Incidentes de Ciberseguridad, que si bien no remite principalmente a la IA, contempla algunas de las potenciales intervenciones de sistemas de IA.

En tanto, el 12 de junio de 2023 se publicó en las páginas del portal oficial del Estado argentino que el Banco Interamericano de Desarrollo (BID) había aprobado un préstamo a cinco años por un valor de 35 millones de dólares destinado al Programa de Apoyo a las Exportaciones de la Economía del Conocimiento con el fin de apoyar el desarrollo del sector y su inserción internacional (Argentina, 2023). Fueron designados como organismos ejecutores el Ministerio de Economía y el Ministerio de Ciencia, Tecnología e Innovación a través de la Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación (Agencia I+D+i).

Por su parte, en la reciente Reforma Constitucional de la provincia de Jujuy, publicada en el Boletín Oficial provincial el 21 de junio de 2023, existe un artículo dedicado especialmente a la IA (el número 76), con el que se reconoce en dicha provincia "el derecho de toda persona a utilizar sistemas de inteligencia artificial o no humana, basados en métodos computarizados de

algoritmos, datos y modelos que imitan el comportamiento humano y automatizan procesos complejos, así como otros futuros desarrollos que surjan en este campo". Y se afirma que se sujetarán esos sistemas "a los principios de legalidad, transparencia, responsabilidad, privacidad y protección de datos, seguridad, no discriminación y rendición de cuentas, garantizando el acceso a la justicia en caso de vulneración de derechos y consagrando la acción de solicitud de revisión humana cuando sea necesario". (5)

En parte por la propia novedad del fenómeno, que suscitó un gran interés en los medios, la conversación pública que lo tiene en el foco se ha caracterizado por no estar del todo informada y por incluir una sobreabundancia de material no sistematizado. A esto se agrega el hecho de que la pregunta por la IA implica distintas dimensiones que, sin el suficiente trabajo de conceptualización, se mantienen plegadas: ¿qué son exactamente estas *metatecnologías*, en qué escalas actúan?, ¿de qué son capaces, cómo funcionan?, ¿en qué se distinguen de tecnologías anteriores?, ¿cuáles son los beneficios y cuáles los peligros potenciales de su aplicación?, ¿qué tipo de diálogo pueden establecer los seres humanos con ellas?, ¿nuestra especie se encuentra en peligro de extinción?, ¿qué papel pueden y deben asumir frente a ellas los Estados Nacionales?

En este artículo presentamos la perspectiva analítica que desarrollamos para analizar este fenómeno. Desde el comienzo entendimos que, para abordar el análisis de los riesgos y desafíos de las IA generativas, era preciso tomar una serie de decisiones que, una vez asumidas, constituyeron parte relevante de la delimitación epistemológica y metodológica de nuestra investigación. Identificamos así cinco rasgos o aspectos de las IA; luego, recortamos las distintas escalas en las que actualmente se desarrollan; y, en tercer lugar, presentamos y distinguimos dos enfoques transversales para su abordaje y tratamiento.

Los principales rasgos de las IA

El primer aspecto a tener en cuenta es que las Inteligencias Artificiales no son *una* tecnología, sino que se trata de *metatecnologías*, esto es, tecnologías de propósito general, aplicables a muy diversas actividades. Y esto es así en al menos tres sentidos.

Por un lado, y principalmente, porque —como escribe Ariel Vercelli en su artículo "Las inteligencias artificiales y sus regulaciones: Pasos iniciales en la Argentina, aspectos analíticos y defensa de los intereses nacionales"—, las IA

> pueden ser analizadas como redes heterogéneas, híbridos y ensambles tecnológicos. Su mera existencia evidencia la articulación e integración de éstas con otras redes, prácticas y procesos científico-tecnológicos más amplios. Al igual que ocurrió con el software (los programa de computación) en las etapas tempranas de la computación electrónica digital, en muchas ocasiones las IA también están indiferenciadas de los dispositivos y sistemas tecnológicos donde están incorporadas (2023, 208).

El hecho de que las IA estén indiferenciadas de los dispositivos y sistemas tecnológicos en los que están incorporadas tiene efectos tanto en el nivel analítico como en el de la gobernanza. Por ejemplo, a los fines regulatorios, no alcanza con establecer una norma general para las IA, sino que es preciso identificar las capas y subsistemas que participan en las IA para analizar las diferentes legislaciones que las atraviesan, desde protección de datos y derechos de autor hasta legislación laboral y de protección del medioambiente.

Por otro lado, tal como señala el filósofo Luciano Floridi, conocido por su trabajo en ética de la IA, se denominan metatecnologías a aquellas tecnologías que "operan y regulan otras tecnologías" (Floridi, 2011, 91). Y si bien no siempre es así, en ocasiones las IA son metatecnologías como las leyes o las tecnologías de seguridad, porque son "parte de las condiciones de operación de otras tecnologías" (ibíd.).

Finalmente, porque tal como afirman Ajay Agrawal, John McHale y Alex Oettl en su artículo "Finding needles in haystacks: Artificial Intelligence and recombinant growth [Encontrar agujas en pajares. Inteligencia artificial y crecimiento recombinante]", las IA también nos ayudan a producir conocimientos que jamás tendríamos sin ellas. En este sentido, la reciente explosión en la disponibilidad de datos y los avances informáticos en las capacidades para descubrir y procesar esos datos "pueden ser vistos como 'metatecnologías', esto es: tecnologías para la producción de nuevos conocimientos" (2018: 3). "Por supuesto", agregan,

> las metatecnologías que ayudan en el descubrimiento de nuevos conocimientos no son nada nuevo. [...] [Pero] la promesa de la IA como metatecnología para la producción de nuevas ideas es que facilita la búsqueda en espacios de conocimiento complejos, permitiendo tanto un mejor acceso al conocimiento relevante como a una mejor capacidad para predecir el valor de nuevas combinaciones (ibíd.).

Es en este mismo sentido que Alex Trollip (2021) escribe que "las metatecnologías son tecnologías o invenciones que tienen la capacidad de ayudar a nuevos descubrimientos o estimular la innovación en otras áreas".

El segundo rasgo a considerar es que las inteligencias artificiales, en la medida en que integran y expanden el ecosistema digital, constituyen no una herramienta o dispositivo técnico, sino que –particularmente después del shock de virtualización que implicó la pandemia de coronavirus (Costa, 2021)– han comenzado a ser para nosotros un mundoambiente. Las tecnologías del ecosistema digital están dejando de ser instrumentos que podemos elegir usar o no, y se han vuelto cada vez más indispensables para realizar actividades cotidianas como guiarnos en una ciudad o iniciar un trámite de documentación obligatorio (Zuboff, 2020). Esto significa que, en relación con los usuarios, no es suficiente un enfoque que las aborde desde la perspectiva de la instrumentalidad y de la relación sistema-usuario individual, como si cada usuario pudiera decidir qué hacer en cada caso con la IA, sino que es necesario un enfoque sistémico, atento a las dinámicas multiescalares de la economía política del ecosistema digital.

El tercer rasgo es particularmente significativo para los estudios sociales y las humanidades. Consiste en señalar que, a los efectos de un análisis epistemológico con epicentro en las ciencias sociales, las llamadas IA generativas y en particular los Modelos de Lenguaje Grandes (LLM, por sus siglas en inglés) son, no sólo Inteligencia Artificial, sino también Sociedad Artificial. Esto es así debido a que los LLM operan desde y sobre el mundo social a través del sistema de Datos, Algoritmos y Plataformas (DAP). (6)

Veamos esto. Los elementos básicos de los LLM son tres: (1) enorme capacidad de cómputo (a grandes rasgos, la capa del hardware), (2) métodos de procesamiento de información (aprendizaje profundo, redes neuronales, etc.; las capas del software y de las aplicaciones de IA) y (3) grandes conjuntos de datos (materiales "de la vida social", obtenidos en buena medida a través de plataformas; las capas de *input* y de usuarios). Es decir: su

alimento (*input*) y su producción específica (*output*) son los intercambios lingüísticos en diferentes idiomas, las figuras retóricas y las reacciones emocionales, las relaciones sociales de diversas culturas. Estos sistemas sociotécnicos complejos que son los LLM y las IA generativas aceleran el procesamiento, la gestión y la (re)producción de lo social. *Producen sociedad*. Y si las capas 1 y 2 son el producto de desarrollos históricos de las ciencias de la informática y la computación, el estudio y el trato con la capa 3 es el dominio de las ciencias del lenguaje, las ciencias de la comunicación, la sociología y la ciencia política, entre otras disciplinas. De allí que es deseable que en su desarrollo, su análisis y su monitoreo participen expertos en esos campos disciplinares, que a su vez deberán estar formados y entrenados en el trato con estos sistemas técnicos.

El cuarto rasgo consiste en que, en ciertos usos, las IA pueden ser tecnologías de alto riesgo, y por lo tanto requieren de un tratamiento acorde a lo largo de todo su ciclo de vida. Es preciso distinguir cuáles son esos casos o usos, y abordarlos con la perspectiva sistémica de la seguridad y la gestión de riesgos. Esto es lo que hace la Ley de Inteligencia Artificial de la Unión Europea, que distingue diferentes tipos de usos y prácticas en los que puedan participar sistemas de IA y establece tres categorías de riesgo. La primera es la de los riesgos inaceptables, lo que significa que hay usos de IA que están prohibidos. Por ejemplo, la manipulación maliciosa del comportamiento, la calificación social y la vigilancia masiva. Otros riesgos se consideran altos y deben ser obligatoriamente gestionados; esto implica un complejo circuito de gestión de calidad y riesgo, documentación, certificaciones y notificaciones. Ejemplos de esta segunda categoría son la identificación biométrica y la categorización de personas; la gestión y el funcionamiento de infraestructuras esenciales; la educación y la formación profesional; el empleo, la gestión de los trabajadores y el acceso al autoempleo; el acceso a servicios esenciales; y la aplicación de la ley. Una tercera categoría es la de los riesgos mínimos, en los que la UE exige transparencia para con el usuario: desarrolladores e implementadores deben informar al usuario que está interactuando con un sistema de IA. La ley europea es exhaustiva al describir la red institucional a cargo de la gestión de esos riesgos, e incluye la obligación de informar accidentes o incidentes de IA en no más de 72 horas desde su ocurrencia a las autoridades nacionales y a las autoridades de la UE. Por su parte, la OCDE lanzó en noviembre de 2023 el Monitor de Incidentes de IA (AIM), desarrollado por el grupo de expertos en incidentes de IA que, en paralelo, está trabajando en un marco de notificación de

incidentes de IA. El objetivo deL AIM es realizar un seguimiento de los incidentes de IA en tiempo real y proporcionar la base de evidencia para informar el marco de notificación de incidentes de IA y los debates sobre políticas de IA relacionados.

El quinto elemento no es tanto un rasgo de las IA generativas sino una consecuencia analítica de tener en cuenta los rasgos anteriores. Para afrontar las IA generativas desde las ciencias sociales y humanas, para pensar eficazmente su gobernanza, no es suficiente la perspectiva de la ética de la IA, que es la forma más habitual en la que los saberes de las ciencias humanas se presentan en la discusión (por ejemplo, en la *Recomendación sobre la ética de la inteligencia artificial* de la Unesco, adoptada por los países miembro en 2021 [Unesco, 2022]). Si de lo que estamos hablando es de metatecnologías que constituyen un mundoambiente, que aceleran el procesamiento y la gestión de lo social, y que pueden ser de alto riesgo en áreas de experiencia críticas para la población como el acceso a la salud, al empleo Y a la educación, nos encontramos ante sistemas sociotécnicos complejos. Por ende, la perspectiva que necesitamos va más allá de las recomendaciones voluntarias, que ofician como “códigos de buenas prácticas”.

Como dice Brent Mittelstadt en su artículo “Principles alone cannot guarantee ethical AI” , no alcanza con establecer principios: “Se necesitan estructuras de rendición de cuentas vinculantes y altamente visibles, así como procesos claros de implementación y revisión a nivel sectorial y organizacional” (2019: 4).

Hasta ahora, las iniciativas de ética de la IA han producido principios y declaraciones de valores que prometen guiar la acción, pero en la práctica brindan pocas recomendaciones específicas y no abordan normas y políticas fundamentales. Siempre según Mittelstadt, entre las tareas necesarias una de ellas es licenciar a los desarrolladores de IA de alto riesgo, definiendo requisitos claros de confiabilidad y reputación. Señala el autor:

> Puede ser necesario establecer el desarrollo de la IA como una profesión con una categoría equivalente a otras profesiones de alto riesgo –asegura–. Es una rareza regulatoria que otorguemos licencias a profesiones que brindan un servicio público, pero no a la profesión responsable de desarrollar sistemas técnicos para aumentar o reemplazar la experiencia humana y la toma de decisiones dentro de ellos. Los riesgos de las profesiones autorizadas no se han disipado, sino que han sido desplazados a la

> IA. Para analizar los importantes desafíos que enfrentan, las iniciativas podrían dirigirse inicialmente a los desarrolladores de diseño inclusivo, revisión ética transparente, documentación de modelos y conjuntos de datos, y auditoría ética independiente (ibíd., 9-10).

De manera afín, Vercelli señala la necesidad de no confundir "los problemas éticos de las IA con las políticas y las regulaciones", ya que "¿qué poder jurídico-político tienen las recomendaciones sobre ética de la IA de la Unesco? Siempre son útiles, pero se trata sólo de meras recomendaciones". Para el investigador argentino, "es necesario comprender que estas posiciones éticas podrían no ser el mejor enfoque para avanzar sobre políticas públicas y regulaciones nacionales". En cambio, sugiere "superar los comités de ética y ofrecer a la sociedad una discusión más amplia, abierta y democrática sobre IA". Desde su perspectiva, el peligro de las IA es que "es que profundicen las injusticias del mundo real: desigualdades sociales, económicas, jurídico-políticas y ambientales". De allí que se requieren procesos de co-construcción entre tecnologías y regulaciones; esto es, "desarrollar tecnologías (incluso para fines regulativos: control y gestión del tiempo, el espacio y las conductas)" (2023, 213-214)

Para afrontar las IA generativas desde las ciencias sociales y humanas, para que su gobernanza sea eficaz, no es suficiente con un enfoque desde la ética profesional de la IA, sino que es preciso promover un enfoque desde la ética organizacional de la IA y desde el pensamiento sistémico, que busca establecer procedimientos de revisión transparentes, estructuras de rendición de cuentas vinculantes, documentación de modelos y conjuntos de datos, auditoría independiente, etc. En suma: defensas en profundidad a lo largo del sistema para que este sea más seguro y confiable.

Las tres escalas de la IA

Con el doble propósito de ordenar la conversación pública y, al mismo tiempo, identificar el ámbito más propicio para el despliegue de las iniciativas estatales, proponemos distinguir entre tres escalas de la IA: una escala macro, una escala meso y una escala micro. Cada una de

ellas comporta actores, riesgos y desarrollos tecnológicos específicos. Veamos el siguiente cuadro, que es una primera aproximación todavía en proceso.

Escala	Macro	Meso	Micro
Actores paradigmáticos	Corporaciones info-tecnológicas (como Open IA, Alphabet, Meta y Google)	Estados (nacionales, provinciales, municipales) e instituciones interestatales	Actores privados (como industrias pequeñas y medianas) o estatales de escala local
Riesgos identificados	Riesgo existencial, armamentismo, debilitamiento de las capacidades humanas, desinformación, manipulación del comportamiento, securitización del conocimiento, erosión epistémica	Multiplicación de sesgos, datos inadecuados o malinterpretados, suplantación de identidad, desinformación, vigilancia, manipulación del comportamiento, securitización del conocimiento	Datos inadecuados mal interpretados, opacidad (exigencia de transparencia)
Desarrollos tecnológicos: ejemplos	Modelos de lenguaje grandes (LLM), como el Chat GPT; algoritmos de redes sociales	Detección de fraudes en la asistencia social (SyRI), sistema de redacción de fallos judiciales (Prometea), GPS y otros sistemas de geolocalización	Dispositivos de visión 3D, reconocimiento de patrones en imágenes para aplicaciones de salud, monitores inteligentes para pacientes que están en cuidados intensivos,

			optimización de láseres para metrología

Figura Nº 1. “Las tres escalas de la IA”. Fuente: elaboración propia.

Escala macro

Los actores que participan de la escala macro del desarrollo de la IA son relativamente pocos. Se trata de *Big Tech* tales como Open AI, Google, Alphabet, Meta, Amazon, DeepMind, Yandex, Huawei, Laion o Baidu. Estas empresas comparten una serie de características principales: acceden a enormes volúmenes de datos; tienen desarrollos de vanguardia en materia de métodos y capacidad de procesamiento; y, aunque tengan anclaje nacional, operan a nivel transnacional. Es en esta escala que se han desarrollado tecnologías como los modelos de lenguaje grandes como el que soporta al famoso Chat GPT, de Open AI.

En este nivel de alto desarrollo de la IA están actualmente en debate una serie de problemas específicos. Por ejemplo, la especulación en torno a las consecuencias que podría desencadenar la emergencia de una superinteligencia artificial, en los términos del filósofo sueco Nick Bostrom (2016), o de una “singularidad”, como la llama el inventor y empresario transhumanista Raymond Kurzweil.

En su libro *La singularidad está cerca* (2012), Kurzweil, desde 2012 director de ingeniería de Google, afirma que la Inteligencia Artificial hará en las próximas décadas un salto de escala. Según su tesis, primero pasará de ser la inteligencia artificial estrecha que conocemos hoy (especializada en una sola tarea, como guiarnos en una ciudad) a ser una inteligencia artificial general, que será al menos tan desarrollada como un ser humano en distintos ámbitos (ese sería acaso el momento que ahora mismo estamos atravesando). Y que de allí saltaría a una superinteligencia artificial, mucho más veloz y más inteligente que cualquier humano, e incluso que la humanidad en su conjunto. Considera imposible imaginar el futuro humano después de ese punto de inflexión:

> La singularidad constituirá la culminación de la fusión entre nuestra existencia y pensamiento biológico con nuestra tecnología, dando lugar a un mundo que seguirá siendo humano pero que trascenderá nuestras raíces biológicas. En la post-singularidad, no habrá distinción entre humano y máquina o entre realidad física y virtual (Kurzweil, 2012, 9-10).

Böstrom, en tanto, define a la superinteligencia "como cualquier intelecto que exceda en gran medida el desempeño cognitivo de los humanos en prácticamente todas las áreas de interés" (2016: 52). De acuerdo con este autor, uno de los fundadores de la Asociación Transhumanista Mundial, el desarrollo de la superinteligencia en el marco de un modelo competitivo como el que actualmente guía el desarrollo de estas tecnologías podría ser muy peligroso. (7)
Dice Bostrom respecto del riesgo existencial:

> la primera superinteligencia podría dar forma al futuro de la vida de origen terrestre, podría fácilmente tener objetivos finales no antropomórficos, y, probablemente, tendría razones instrumentales para perseguir la adquisición indefinida de recursos. Si ahora reconocemos que los seres humanos constituyen recursos útiles (como átomos convenientemente ubicados) y que dependemos para nuestra supervivencia y nuestra realización de muchos más recursos locales, podemos ver que el resultado podría ser fácilmente uno en el que la humanidad fuera rápidamente extinguida" (ibíd., 171).

De esta manera, para Böstrom, la lógica de la competencia podría tener como paradójico resultado final la no competencia: una vez que la lucha por la competitividad escala por fuera de nuestro control, la ventaja competitiva de esa primera superinteligencia podría ser el lograr que no exista ningún adversario. A este tipo de peligro se suele aludir habitualmente con la noción de "riesgo existencial", al que se reconoce como un caso particularmente grave de riesgo de las IA.

En el campo de la investigación en IA, la indagación sobre la alineación o el alineamiento (en inglés, *AI alignment*) se ocupa de dirigir el desarrollo de los sistemas de inteligencia artificial de acuerdo con los objetivos e intereses de sus diseñadores. Si un sistema es eficiente pero persigue objetivos que no han sido previstos por los investigadores, se dice

que no está alineado. La alineación de los sistemas de inteligencia artificial incluye desafíos como la dificultad de especificar completamente todos los comportamientos deseados y no deseados; el uso de objetivos intermedios fáciles de especificar que omiten restricciones deseables; objetivos instrumentales, como la búsqueda de poder, que ayudan al sistema a lograr sus objetivos finales o emergentes pero que sólo se vuelven evidentes cuando el sistema se implementa en nuevas situaciones y distribuciones de datos. Estos problemas afectan tanto a robots como a modelos de lenguaje, vehículos autónomos y sistemas de recomendación de redes sociales. Se cree que los problemas de alineamiento son más probables cuanto más potente es el sistema, ya que en parte resultan de esa enorme potencia. De allí que la comunidad de investigadores de la inteligencia artificial y organizaciones como las Naciones Unidas o la OCDE han señalado la necesidad de impulsar soluciones tanto técnicas como políticas para garantizar que los sistemas estén alineados con los valores humanos.

El Center for AI Safety (s/f) –una asociación sin fines de lucro dedicada a la investigación en seguridad y riesgos de la IA y liderada por el ex OpenIA Dan Hendrycks– ha organizado en cuatro tipos los riesgos asociados a lo que aquí identificamos como escala macro: (a) el uso malicioso (las personas podrían aprovechar intencionalmente potentes IA para causar daños generalizados), (b) la carrera de IA (la competencia podría empujar a naciones y corporaciones a apresurar el desarrollo de la IA, cediendo el control a estos sistemas; las corporaciones estarían incentivadas a automatizar el trabajo humano, lo que podría conducir a desempleo masivo y dependencia de los sistemas de IA; además, a medida que proliferan los sistemas de IA, la dinámica evolutiva sugiere que serán más difíciles de controlar); (c) los riesgos organizacionales (particularmente si las organizaciones priorizan las ganancias sobre la seguridad); y (d) las IA rebeldes (podríamos perder el control sobre las IA a medida que se vuelvan más capaces).

Respectivamente para cada caso, ellos recomiendan (a) mejorar la bioseguridad, restringir el acceso a modelos de IA peligrosos y responsabilizar a los desarrolladores de IA por los daños. (b) Desarrollar normas de seguridad, coordinación internacional y control público de las IA de uso general. (c) Fomentar una cultura organizacional orientada a la seguridad e implementar auditorías rigurosas, defensas contra riesgos de múltiples niveles y seguridad de la información de última generación. Y (d) Impedir que las IA se implementen en entornos de alto riesgo, como la búsqueda autónoma de objetivos abiertos o la supervisión de

infraestructuras críticas, a menos que se demuestre que son seguras. También avanzar en la investigación de la seguridad de la IA en áreas como la robustez frente a ataques adversarios, la transparencia y la eliminación de capacidades no deseadas. (8)

Escala micro

Al otro extremo de la escala macro, la escala micro comprende los desarrollos específicos que se implementan en muy diferentes industrias y disciplinas y que, en principio, no parecen requerir de un cuerpo de regulación específico. Entre los muchísimos ejemplos que pueden ofrecerse, cabe citar desde sistemas de optimización de láseres para metrología hasta el monitoreo de pacientes en cuidados intensivos, entre muchos otros usos de la IA que pueden realizarse en pequeñas y medianas industrias, o en dependencias de nivel estatal o privado, para mejorar los resultados de alguna tarea concreta.

Escala meso

En el marco de nuestra investigación, orientada a efectuar recomendaciones para el ámbito público, identificamos la escala meso como aquella más específica para situar las políticas públicas en los niveles nacional, provincial y municipal –si bien los riesgos de la escala macro no pueden ser negados, en relación con la escala de desarrollo nacional, constituyen un marco general a tener en cuenta, pero sobre el cual en principio se actúa sólo de manera indirecta–.

Algunos de los riesgos más habitualmente señalados por la literatura internacional (HM Government, 2023; OECD, 2023) en relación con esta escala son los sesgos; los datos inadecuados o malinterpretados; la suplantación de identidad; la desinformación (*deep fake*, ya provenga de humanos o de máquinas); la vigilancia; la manipulación del comportamiento; y, por último, la securitización del conocimiento, es decir, el hecho de que el conocimiento experto queda en manos de cada vez menos personas. Se trata, en conjunto, de riesgos para la calidad democrática que están siendo investigados desde perspectivas y disciplinas no siempre suficientemente comunicadas entre sí.

Un ejemplo de incidente de IA en esta escala lo proporciona lo ocurrido con el software System Risk Indication (SyRI), utilizado entre 2014 y 2019 por el gobierno de los Países Bajos

para detectar fraudes en el pedido de asistencia social. En 2018, seis organizaciones no gubernamentales presentaron una demanda contra el sistema, que ya había dejado a cerca de 10 mil familias en problemas, debido a que no solo le quitaron la ayuda, sino que debieron devolver la recibida hasta el momento. A principios de 2020, los tribunales dictaminaron que el sistema SyRI no había logrado un equilibrio justo entre detección de fraude y privacidad. Sostuvieron que el sistema SyRI era demasiado opaco, recopilaba demasiados datos y los propósitos para recopilarlos no eran lo suficientemente claros y específicos. Este fue uno de los primeros casos en Europa que cuestionó el uso estatal de software de calificación de riesgos de "vigilancia policial predictiva". Al año siguiente, la plana mayor de la coalición en el gobierno debió renunciar por este escándalo.

Reaparece así la preocupación por el alineamiento de los sistemas de IA –y en particular aquellos sistemas que se identifican como de alto riesgo, en tanto participan en procedimientos que pueden afectar derechos fundamentales tales como el acceso a la justicia, a la salud, a la educación– con valores sociales y humanos, por cómo incluir en el diseño de estos sistemas técnicos el cuidado de la vida, del medio ambiente, de la especie, etcétera. La protección de valores fundamentales que, si no son incluidos en su programación, no se verán luego reflejados en su operación.

Es en esta escala que intentan operar las leyes, normas, pautas y recomendaciones para el desarrollo y la implementación de IA *confiables* y *seguras*. Los objetivos generales que se persiguen a través de la idea de "fiabilidad" o "confiabilidad" de la IA son cuatro: la robustez de los desarrollos y las implementaciones de IA, lo cual incluye trabajar con datos de buena calidad, construidos con meticulosidad, resguardando que no repliquen sesgos; el monitoreo del desempeño de los sistemas de IA tanto desde el diseño como durante el desarrollo y la implementación, debido a que la deriva práctica del sistema siempre puede ser diferente de su performance proyectada; la alineación, que supone que los sistemas de IA tengan mecanismos de reaseguro respecto de la alineación con los valores humanos; y como modelo epistémico, un pensamiento sistémico de la seguridad, capaz de abordar de manera integral sistemas sociotécnicos complejos como las IA.

Dos enfoques de la IA

Algo que hemos observado en nuestra investigación sobre “Desafíos e impactos de la inteligencia artificial” (2023) es que los expertos que provienen de las ciencias de la computación, la informática y las ingenierías suelen identificar como su principal (a veces, el único) interlocutor en el campo de las ciencias sociales al derecho. Es su punto de exterioridad y su límite: el establecimiento de normas que regulen la práctica. Y si se traspasa el límite establecido, hay una sanción. Pero el derecho actúa *a posteriori* del accidente o incidente. Y si entendemos que en algunos casos las tecnologías de IA pueden ser de alto riesgo, la aseguración (*safety*) debe ser proactiva, preventiva y predictiva; debe estar inscripta en el propio sistema. En los sistemas sociotécnicos complejos, la evaluación de riesgos y el monitoreo de seguridad es parte de todo el ciclo de vida del sistema. Con todo, dado que la seguridad de la IA es un campo emergente en el que hay poca casuística, será importante seguir de cerca el proceso que está llevando adelante el monitor de incidentes de IA que recientemente puso en marcha la OCDE.

De allí que, en este apartado, nos interesa distinguir dos grandes enfoques transversales y complementarios frente a los desafíos que implica la IA: el enfoque jurídico, orientado a la responsabilidad, y el enfoque sistémico, orientado a la protección (*security*) y a la seguridad (*safety*).

Estos enfoques se distinguen entre sí por sus objetivos: mientras que el primero se propone atribuir las responsabilidades ante un fallo o accidente, procurando encontrar la causa del incidente y proponiendo algún tipo de sanción, pena o resarcimiento en el caso de que corresponda, el segundo apunta a mejorar la *performance* del sistema, identificando los riesgos asociados a determinada actividad, intentando controlarlos y que se mantengan en un nivel aceptable.

Este último enfoque es el propio de las industrias de alto riesgo, que apuntan a ser ultra seguras (Covello, 2021). (9) Desde nuestro punto de vista, la IA debe ser tratada en esta clave, entendiendo que un enfoque basado en el riesgo resulta adecuado para asegurar que la intervención gubernamental sea proporcionada. Esto es: que contribuya eficazmente a mitigar los riesgos de impactos negativos, pero que no sea excesivamente prescriptiva ni obstaculice el desarrollo y la implementación de IA en campos en los que su aporte puede ser especialmente benéfico.

Enfoque	Jurídico	Pensamiento sistémico de la seguridad
Instrumentos	Leyes, normas, regulaciones, reglamentos Auditoría Investigación judicial	Gestión de riesgos y Garantía de la seguridad Monitoreo, auditoría Investigación de accidentes e incidentes
Intervenciones	Penalidades, sanciones resarcimientos	Recomendaciones, ajustes del sistema
Objetivos	Acceso a la Justicia y aplicación de la ley	Robustez del sistema

Figura Nº 2. “Dos enfoques de la IA”. Fuente: elaboración propia

Palabras finales

Este artículo describe elementos clave de la perspectiva analítica y enumera un conjunto de decisiones teórico-epistemológicas desarrollados en 2023 en el marco de una investigación sobre los desafíos y riesgos de las IA. Entre las recomendaciones que emergen de esta investigación, entendemos que es importante para nuestro país desarrollar un marco normativo para las IA, para el cual sugerimos un enfoque basado en el riesgo. Este resulta particularmente equilibrado para, por un lado, contribuir eficazmente a mitigar los riesgos de

impactos negativos, y por otro, no obstaculizar el desarrollo y la implementación de IA en campos en que su aporte puede ser especialmente beneficioso.

Asimismo, entendemos que es necesario impulsar una iniciativa nacional coordinada de desarrollo y gobernanza de las, que IA debe plantearse al menos seis objetivos: generar capacidades de dirección y gestión de proyectos multidisciplinarios de desarrollo y monitoreo de tecnologías basadas en IA; articular las capacidades del sistema científico y tecnológico en IA y las necesidades del sector productivo; elaborar una propuesta de política regulatoria de IA con perspectiva jurídica y enfoque sistémico; promover la formación de expertos "bilingües" capaces de comprender los desafíos sociales, políticos y educativos de la IA; promover la formación de expertos en gestión de riesgos e investigación de incidentes de IA; y contribuir a la internacionalización del ecosistema de IA.

Notas

(1) Este trabajo se nutre en parte de la investigación "Desafíos e impactos de la Inteligencia Artificial. Marcos normativos, riesgos y retos para la calidad democrática en la Argentina", realizada por las y los autores en la Facultad de Ciencias Sociales de la Universidad de Buenos Aires a solicitud del Programa "Argentina Futura", dependiente de la Jefatura de Gabinete de Ministros de la Nación, entre julio y noviembre de 2023.

(2) La OCDE (2019) define a las IA como sistemas basados en máquinas que pueden, dado un conjunto determinado de objetivos definidos por el ser humano, realizar predicciones y recomendaciones o tomar decisiones que influyen en entornos reales o virtuales. Señala, además, que los sistemas de IA están diseñados para funcionar con diversos niveles de autonomía.

(3) En junio de 2023, la UE discutió y dio por terminado el proceso de enmiendas a la Ley de IA que regirá por los próximos años. El 14 de ese mes, la plenaria del Parlamento Europeo aprobó un proyecto con enmiendas a la Ley de 2021 para regular el uso de la Inteligencia Artificial en la Unión Europea, dando inicio a una delicada negociación con los 27 países del bloque. La normativa sancionada (con 499 votos a favor, 28 en contra y 93 abstenciones) ratifica la regulación de la IA según el nivel de riesgo: cuanto mayor sea este para los derechos

o la salud de las personas, mayores serán las obligaciones de los sistemas tecnológicos. La perspectiva de los Estados Unidos, en cambio, es diferente. En este país, la principal norma relacionada con los sistemas de IA no adopta principalmente la perspectiva del riesgo, sino que busca asegurar el liderazgo mundial de los Estados Unidos en relación con la investigación y el desarrollo de sistemas de IA. Dicha norma se denomina Iniciativa Nacional de Inteligencia Artificial (*National Artificial Intelligence Initiative Act*, *NAII*), fue sancionada en 2020 y entró en vigencia el 1° de enero de 2021.

(4) Vale señalar que existió una iniciativa nacional anterior, elaborada durante la presidencia de Mauricio Macri entre 2018 y 2019, el Plan Nacional de Inteligencia Artificial (ArgenIA, 2019), que señala la importancia de las IA, afirma la necesidad de formar recursos humanos, la relevancia de utilizar datos de calidad, pone el acento en la infraestructura computacional, y propone la creación de un Laboratorio de Innovación para "acelerar y canalizar el cumplimiento de objetivos propuestos en el Plan Nacional de IA". Con el cambio de gobierno y la irrupción al año siguiente de la pandemia mundial del COVID-19, el Plan quedó sin efecto. Es posible consultarlo en Internet: ia-latam.com/wp-content/uploads/2020/09/Plan-Nacional-de-Inteligencia-Artificial.pdf

(5) El texto puede consultarse en saij.gob.ar/0-local-jujuy-constitucion-provincia-jujuy-lpy0000000-1986-10-22/123456789-0abc-defg-000-0000yvorpyel

(6) Los Modelos de Lenguaje Grandes o Grandes Modelos de Lenguaje (*Large Language Models*, LLM) son un ejemplo de IA generativa que produce textos o código, que se han popularizado recientemente por la difusión y el uso creciente de herramientas como el mencionado Chat GPT. Se trata de un tipo de Modelo Básico o Fundacional, que opera con algoritmos basados en redes neuronales artificiales entrenadas con inmensos conjuntos de datos sin etiquetar, autosupervisados para producir texto y código significativo de manera similar a los que puede crear un ser humano. Los LLM pueden generar textos fluidos y coherentes sobre diversos temas (Bowman, 2023; Keen, 2023).

(7) Para comprender su argumento, vale la pena recuperar una noción propia del ámbito de la filosofía de la técnica propuesta por Andrew Feenberg (2002), la de "código técnico". Esto es, el tipo de orientación histórica que recibe el dominio de la actividad técnica. De acuerdo con

Feenberg, el código técnico capitalista es competitivo y condiciona el desarrollo de los conjuntos técnicos en esa dirección.

(8) Para profundizar en esta línea véase Hendrycks, Mazeika y Woodside (2023).

(9) En nuestro país, la ley N° 27514 de creación de la Junta de Seguridad en el Transporte (JST) define a la seguridad operacional como el estado de operación de un sistema en que el riesgo de lesiones a personas o daños a los bienes que participan e interactúan, se ve reducido y se mantiene en un nivel aceptable o por debajo del aceptable, por medio de un proceso continuo de identificación de peligros y gestión de riesgos.

Bibliografía

Agrawal, A.K.; McHale, J.; y Oettl, A. (2018). "Finding needles in haystacks: Artificial Intelligence and recombinant growth". *Nber Working Paper Series*, 24541.

Albrieu, R. *et al.* (2018). *Inteligencia artificial y crecimiento económico. Oportunidades y desafíos para Argentina*, Buenos Aires: Cippec.

Arenas, M.; Arriagada, G.; Mendoza, M.; Prieto, C. (2020). *Una breve mirada al estado actual de la Inteligencia Artificial*, Pontificia Universidad Católica de Chile, 2020. En Internet: https://desarrollodocente.uc.cl/wp-content/uploads/2020/09/Una-breve-mirada-al-estado-actual-de-la-Inteligencia-Artificial.pdf.

Bostrom, N. (2016 [2014]). *Superinteligencia. Caminos, peligros, estrategias*, España: Teell.

Bowman, S.R. (2023). "Eight Things to Know about Large Language Models". *Cornell University*. Recuperado de arxiv.org/abs/2304.00612

Center for AI Safety (s/f). *Center for AI Safety*, "An Overview of Catastrophic AI Risks". Recuperado de safe.ai/ai-risk.

Costa, F. (2021). *Tecnoceno. Algoritmos, biohackers y nuevas formas de vida*, Buenos Aires: Taurus.

Covello, A (2021). *Investigación sistémica de accidentes. Modelo para el transporte y la gestión de riesgo en sistemas complejos*, CABA: Ediciones Ciccus.

Crawford, K. (2022). *Atlas de inteligencia artificial. Poder, política y costos planetarios*, Buenos Aires: Fondo de Cultura Económica.

Feenberg, A.(2002). *Transforming Technology: A critical theory revisited*, Nueva York: Oxford University Press.

Floridi, L. (2011). "Energy, Risks, and Metatechnology". *SRRN*. Recuperado de ssrn.com/abstract=3854445

Gómez Mont, C. *et al.* (2020). *La Inteligencia Artificial al servicio del bien social en América Latina y el Caribe: panorámica regional e instantáneas de doce países*. Banco Interamericano de Desarrollo.

Hendrycks D.; Mazeika M.; y Woodside T. (2023). *An Overview of Catastrophic AI Risks*. s/d. Recuperado de arxiv.org/pdf/2306.12001.pdf

HM Government (2023). *Safety and Security Risks of Generative Artificial Intelligence to 2025*, Reino Unido: Gobierno del Reino Unido.

Keen, M. (2023). How Large Language Models Work [archivo de video]. Recuperado de https://www.youtube.com/watch?v=5sLYAQS9sWQ

Kurzweil, R. ([2005] 2012). *La Singularidad está cerca. Cuando los humanos transcendamos la biología*, Berlín: Lola Books.

Mittelstadt, B. (2019). "Principles alone cannot guarantee ethical AI". *Nat Mach Intell*, 1, 501–507.

OCDE (2019a). *Artificial Intelligence in Society*. doi.org/10.1787/eedfee77-en

OCDE (2019b). *Recommendation of the Council on Artificial Intelligence*. OECD. Recuperado de oecd.ai/en/assets/files/OECD-LEGAL-0449-en.pdf

Subsecretaría de Tecnologías de la Información (2023). *Recomendaciones para una inteligencia artificial fiable*. Buenos Aires.

Trollip, A. (2021). "Unfolding AI's Potential: How Investing in Research and Development Can Produce New Knowledge". *Bipartisan Policy Center*. Recuperado de bipartisanpolicy.org/blog/unfolding-ais-potential/

Unesco (2022). *Recomendación sobre la ética de la inteligencia artificial*. París: Unesco. Recuperado de unesdoc.unesco.org/ark:/48223/pf0000381137_spa

Vercelli, A. (2023). "Las inteligencias artificiales y sus regulaciones: pasos iniciales en Argentina, aspectos analíticos y defensa de los intereses nacionales". *Revista de la Escuela del Cuerpo de Abogados y Abogadas del Estado*, 7(9), 195-217.

Zuboff, S. (2020). *La era del capitalismo de la vigilancia*, Barcelona: Paidós.