OEI

PONTIFICIA COMISIÓN PARA AMÉRICA LATINA

Número especial

Septiembre 2026

# Compromisos éticos frente a la **inteligencia artificial**

REVISTA IBEROAMERICANA DE CIENCIA, TECNOLOGÍA Y SOCIEDAD

ISSN:1850-0013

# REVISTA IBEROAMERICANA DE CIENCIA, TECNOLOGÍA Y SOCIEDAD

## Número especial: “Compromisos éticos frente a la inteligencia artificial”

**REVISTA IBEROAMERICANA DE CIENCIA, TECNOLOGÍA Y SOCIEDAD**

Edición especial: "Compromisos éticos frente a la inteligencia artificial"
Septiembre de 2026
Coordinación editorial: Luis Scasso

ISSN: 1668-0030 - ISSN online: 1850-0013

Diseño y diagramación: Julián Fernández Mouján

**CONTACTO**



Observatorio Iberoamericano de la Ciencia, la Tecnología y la Sociedad (OCTS) de la OEI // Paraguay 1510 (C1061ABD), Buenos Aires, Argentina - Tel./Fax: (54 11) 4813-0033/0034 - Correo electrónico: secretaria@revistacts.net.

*CTS* es una revista académica interinstitucional del campo de los estudios sociales de la ciencia y la tecnología. Publica trabajos originales e inéditos que abordan las relaciones entre ciencia, tecnología y sociedad, desde una perspectiva plural e interdisciplinaria y con una mirada iberoamericana, y es editada por la Organización de Estados Iberoamericanos (OEI) para la Educación, la Ciencia y la Cultura, la Universidad de Salamanca (España), el Centro REDES (Argentina), la Universidad de Campinas (Brasil) —a través de Labjor— y el Instituto Universitario de Lisboa (Portugal). La Secretaría Editorial está a cargo del OCTS de la OEI.

*CTS* está incluida en: Dialnet, EBSCO (Fuente Académica Plus), International Bibliography of the Social Sciences (IBSS), Latindex, Latindex Catálogo 2.0, Red de Revistas Científicas de América Latina y el Caribe (REDALYC), SciELO, Red Iberoamericana de Innovación y Conocimiento Científico (REDIB), European Reference Index for the Humanities and Social Sciences (ERIH PLUS). A su vez, *CTS* forma parte de la colección del Núcleo Básico de Revistas Científicas Argentinas y cuenta con el Sello de Calidad de Revistas Científicas Españolas de la Fundación Española para la Ciencia y la Tecnología (FECYT).

# REVISTA IBEROAMERICANA DE CIENCIA, TECNOLOGÍA Y SOCIEDAD

## Índice

**ARTÍCULOS SELECCIONADOS**



**SOBRE ESTE NÚMERO**

# PRESENTACIÓN

# Compromisos éticos frente a la inteligencia artificial

**Mariano Jabonero ***



La expansión de la inteligencia artificial (IA) es una de las transformaciones más decisivas en la historia de la humanidad, lo cual no implica una mera innovación técnica, sino un fenómeno civilizatorio que ya está redefiniendo la relación que los seres humanos sostenemos con nociones como el conocimiento, el poder, la filosofía y la producción. Si pensamos en términos optimistas, la IA se ofrece a ampliar nuestros horizontes en más de un aspecto. Al dejar en manos de estas nuevas tecnologías las tareas repetitivas y rutinarias que hemos llevado adelante durante siglos, podremos liberar tiempo y energía para desarrollar el pensamiento, el cuidado, la creatividad y la cooperación. Los algoritmos serán capaces de procesar ingentes cantidades de datos que anticiparán crisis medioambientales y sanitarias, eficientizarán la explotación de recursos naturales e individualizan el instrumental educativo para tratar con empatía la trayectoria de cada alumno del futuro. Todo esto podría convertir a la IA en un apéndice con los atributos necesarios para agudizar la percepción del bien común, fortalecer los vínculos sociales y prolongar las herramientas de la razón humana.

Sin embargo, este cúmulo de promesas contiene, a su vez, una serie de desafíos y contradicciones que es imperioso abordar. Existen, ya en nuestro presente, tensiones éticas que podrían trastornar la estructura misma de nuestras sociedades. Una de ellas es la creciente concentración del poder tecnológico en un número reducido de

* Secretario General de la Organización de Estados Iberoamericanos para la Educación, la Ciencia y la Cultura (OEI).

corporaciones y gobiernos que poseen datos masivos e infraestructura para diseñar sistemas avanzados de IA. Esa asimetría impone –o como mínimo amenaza con hacerlo– una nueva forma de desigualdad: la capacidad de decidir qué se considera conocimiento, qué se concibe por verdad y qué tipo de vida merece defensa y protección. En lugar de democratizar la inteligencia, la IA podría terminar consolidando jerarquías rígidas de control y vigilancia a escala planetaria. Por otra parte, a medida que los sistemas adquieren cada vez más autonomía, se desdibuja la frontera de la responsabilidad: ¿quién se hará cargo en las próximas décadas de decisiones tomadas por algoritmos que ni siquiera sus propios creadores comprenden del todo? Este futuro posible trae consigo un límite a la ética tradicional, que se afirma en la imputación de actos conscientes a sujetos individuales, axioma que se torna incompleto ante redes de aprendizaje que actúan sin intención aparente, pero con consecuencias reales. Por último, se debe considerar el riesgo cierto de que la IA –que se alimenta de un gigantesco caudal de datos motorizado por grandes modelos de lenguaje– profundice la replicación de sesgos, prejuicios más o menos dilatados, exclusiones, errores históricos, discriminaciones e iniquidades legitimadas por usos y costumbres nunca desmantelados. Sin malicia, la IA puede reproducir la malicia humana a partir de una opacidad cada vez más compleja de fiscalizar.

Todas estas amenazas han sido descriptas y valoradas en *Magnifica Humanitas*, la primera encíclica del Papa León XIV, publicada en mayo de 2026. En ella, el Sumo Pontífice nos exhorta no solo a regular la IA, sino también a repensar la manera en que habitaremos el mundo junto a ella. El filósofo Byung-Chul Han, al recoger el prestigioso premio Princesa de Asturias de Comunicación y Humanidades 2025, expresó que no estaba en contra de la IA, diciendo a continuación que “puede ser muy útil si se emplea para fines buenos y humanos. Pero también con la inteligencia artificial existe el enorme riesgo de que el ser humano acabe convertido en esclavo de su propia creación. La inteligencia artificial puede ser empleada para manejar, controlar y manipular a las personas. Por eso, la tarea acuciante de la política sería controlar y regular el desarrollo tecnológico de manera soberana”. La educación será, en este contexto, un insumo central; necesitamos empezar a programar con discernimiento, ajustar nuestras expectativas acerca del poder supuestamente infinito de los cálculos y valorar en igual medida aquella zona de la sensibilidad humana que no se visibiliza en datos pasibles de medición. El nuevo debate ético demandará entender el vínculo entre lo humano y lo artificial como una coevolución y no como una guerra de trincheras en la que uno de los lados deberá predominar.



Con ese objetivo, la Organización de Estados Iberoamericanos (OEI) y la Pontificia Comisión para América Latina (PCAL) de la Santa Sede se han propuesto colaborar en una serie de líneas de trabajo vinculadas con la irrupción de la IA en los territorios de la educación, la ciencia y la cultura. Una de estas iniciativas es el número especial de la *Revista Iberoamericana de Ciencia, Tecnología y Sociedad* que hoy presentamos, en el que se conjugan cuestionamientos, reivindicaciones, críticas y propuestas tanto técnicas como morales para arrojar luz sobre la actualidad ética de la IA y su influencia en Iberoamérica, región caracterizada por

desigualdades persistentes y una urgencia cada vez más acuciante por incorporarse de forma integral a una revolución tecnológica que, bien conducida, diversificará las economías nacionales, sofisticará la salud, ahondará la educación y estará en condiciones de resolver problemas de larga data.

Surgidos de una convocatoria abierta y de invitaciones extendidas a distintos expertos, los artículos que nutren el índice acercan contribuciones de investigadores e investigadoras de España, Argentina, México, Costa Rica, Venezuela y Chile. Todos estos textos buscan desplegar caminos para que la IA amplíe las capacidades que nos signan como seres humanos. Si la humanidad logra orientar esta fuerza inédita hacia el bien común, si hace de la IA un espejo crítico y no un sucedáneo de su propia conciencia, entonces el futuro no será una rendición, sino una oportunidad para reinventar lo humano desde la posibilidad de pensar, crear y convivir con cada vez mayor responsabilidad.

# Responsabilidad y discernimiento para alcanzar el bien común

**Emilce Cuda ***

Este número especial sobre inteligencia artificial y ética de la *Revista Iberoamericana de Ciencia, Tecnología y Sociedad* es fruto de una colaboración conjunta entre la Organización de Estados Iberoamericanos (OEI) para la Educación, la Ciencia y la Cultura, y la Pontificia Comisión para América Latina (PCAL). En las siguientes páginas se reúnen diversos artículos académicos, elaborados por profesionales y expertos, que nos invitan a reflexionar sobre la dimensión ética de la inteligencia artificial en el debate contemporáneo. Los autores abordan este desafío desde múltiples perspectivas: políticas y sociales, educativas, públicas y religiosas, incorporando una mirada iberoamericana, que enriquece y diversifica la discusión global.



El Papa León XIV ha hecho suya esta preocupación y ha convocado a la humanidad a participar activamente en la configuración ética y social de esta nueva etapa tecnológica. En sus palabras: “Se trata de responder a otra revolución industrial y a los desarrollos de la inteligencia artificial” (León XIV, 2025). En este sentido, la inteligencia artificial, con su potencial inmenso, requiere responsabilidad y discernimiento para orientar los instrumentos al bien común, de modo que puedan generar beneficios para cada persona y para toda la humanidad.

Como el mismo Pontífice señala en su encíclica *Magnifica Humanitas*, “ahora nos corresponde asumir con lucidez y responsabilidad los retos de nuestro tiempo” (León XIV, 2026, p. 5), siendo preciso “iniciar un discernimiento compartido capaz de profundizar en las raíces espirituales y culturales de las transformaciones que se están produciendo” (León XIV, 2026, p. 6).

* Secretario de la Pontificia Comisión para América Latina (PCAL).

Alentamos que esta publicación contribuya a un diálogo interdisciplinario que ilumine los dilemas éticos de la tecnología desde la perspectiva iberoamericana, inspirada en una visión integral del desarrollo y en el compromiso con un progreso centrado en la dignidad humana, el acceso universal a los bienes tecnológicos, la solidaridad institucionalizada y la lógica de la subsidiariedad, teniendo como punto de partida y llegada a la justicia social en sentido católico (León XIV, 2026, p. 77).

**Bibliografía**

León XIV (2025). Discurso del Santo Padre al Colegio Cardenalicio, 10 de mayo. Ciudad del Vaticano: Santa Sede. Recuperado de: https://www.vatican.va/content/leo-xiv/es/speeches/2025/may/documents/20250510-collegio-cardinalizio.html.

León XIV (2026). Carta Encíclica Magnifica Humanitas, 15 de mayo. Ciudad del Vaticano: Santa Sede. Recuperado de: https://www.vatican.va/content/leo-xiv/es/encyclicals/documents/20260515-magnifica-humanitas.html.

# INTRODUCCIÓN

# Una inteligencia artificial para Iberoamérica

**Ana Capilla ***



Desde hace años la transformación digital, la inteligencia artificial (IA) y otras tecnologías emergentes atraviesan todas nuestras áreas de trabajo en los programas-presupuestos de la Organización de Estados Iberoamericanos para la Educación, la Ciencia y la Cultura (OEI), que aprueban los ministros de Educación de los 23 países iberoamericanos y que rige la actividad de nuestra organización.

En concreto, en la Dirección General de Educación Superior y Ciencia hemos dedicado varias publicaciones específicas a IA, entre las que se destaca el dossier monográfico que acompañó al informe de *El Estado de la Ciencia 2023*.[1] El Observatorio Iberoamericano de la Ciencia, la Tecnología y la Sociedad (OCTS) de la OEI analizó en este dossier la producción científica de la región, y los datos señalan que el número de publicaciones iberoamericanas sobre IA ha aumentado, pero menos que en otras regiones del mundo. Por ese motivo, la participación de Iberoamérica en la producción científica mundial en IA ha descendido, pasando del 9% en 2013 al 5% en 2022.

Estos datos son preocupantes por muchos motivos, entre ellos uno que se apunta en algunos de los artículos que se recogen en este número y que el propio

* Directora de Educación Superior y Ciencia de la Organización de Estados Iberoamericanos para la Educación, la Ciencia y la Cultura (OEI).

[1] La publicación está disponible en: https://oei.int/oficinas/argentina/publicaciones/el-estado-de-la-ciencia-principales-indicadores-de-ciencia-y-tecnologia-iberoamericanos-interamericanos-2023/.

Fondo Monetario Internacional (FMI) ya advirtió en un informe de 2020: la IA puede ampliar la brecha entre las naciones ricas y pobres. Para Iberoamérica, la IA y otras tecnologías como el aprendizaje automático, la robótica, los megadatos y las redes, son fundamentales en la medida en que pueden contribuir a poner fin a una de las trampas de desarrollo de la región: la baja productividad. Para que así sea, es imprescindible que no seamos meros usuarios de algoritmos que otros han diseñado en otras latitudes y en otro idioma. Para un organismo bilingüe como es la OEI, que se dirige a una comunidad de 850 millones de hispanohablantes y lusófonos, es fundamental que podamos desarrollar herramientas de IA que piensen en nuestras dos lenguas.

En la OEI, por tanto, trabajamos a favor de que haya una IA iberoamericana, y esta, necesariamente, tiene que ser una IA ética y justa. Porque tenemos que evitar que el algoritmo perpetúe los prejuicios e inequidades que existen en nuestras sociedades, o que se vulnere nuestra privacidad, nuestros derechos y nuestras libertades fundamentales. A la vez, no podemos desincentivar el trabajo y creatividad de los expertos iberoamericanos que están desarrollando herramientas de IA que pueden ayudar a mejorar nuestra educación o sanidad.

A este dilema, complejo y multifacético, trata de dar respuesta este número especial de la *Revista Iberoamericana de Ciencia, Tecnología y Sociedad*, en el que, a través de una convocatoria abierta llevada adelante en conjunto con la Pontificia Comisión para América Latina (PCAL), hemos invitado a expertos de muy diversas ramas del conocimiento a aportar su contribución. Porque en Iberoamérica necesitamos más IA, pero para que esta sea realmente nuestra, debe de ser respetuosa con unas ciertas premisas éticas.



En realidad, este debate se viene produciendo con cierta periodicidad cada vez que surge una nueva tecnología, especialmente una tan disruptiva como la IA y que, como todo, puede ser susceptible de un uso perjudicial. En esta ocasión es quizás más grave porque la IA toma decisiones que nos afectan directamente; y por ello tenemos que ser firmes al exigir transparencia y claridad en el algoritmo en base al cual se va a tomar esa decisión, así como responsabilidad de quien desarrolla la herramienta de IA y de sus usuarios.

Todas estas reflexiones se recogen en las próximas páginas. Agradecemos a los autores su participación y que hayan querido participar del debate propuesto por la OEI y la PCAL en este número especial, compartiendo con nosotros, y con el público en general, los frutos de su inteligencia humana.

# Pensar éticamente la IA en un mundo en transición

**Luis Scasso** *

Con el propósito central de ubicar a la persona en el centro del desarrollo tecnológico, la primera encíclica del Papa León XIV, *Magnifica Humanitas*, logró en mayo de 2026 un gran impacto en el escenario internacional al reflexionar sobre la necesidad de afirmar y proteger la dignidad humana frente al crecimiento de herramientas digitales como la inteligencia artificial (IA). Las advertencias del Santo Padre tienen un eco especialmente relevante si pensamos en la actualidad (y el futuro) de los países de Iberoamérica, donde la incorporación de estos adelantos coexiste con marcadas heterogeneidades en términos de desarrollo económico y social, así como desafíos en temas como educación, salud y productividad.



La discusión sobre la inteligencia artificial se inscribe, además, en una transformación de mayor alcance. Vivimos una transición civilizatoria caracterizada por la aceleración del cambio tecnológico, la creciente complejidad, la inestabilidad y la incertidumbre. La IA no constituye simplemente una nueva tecnología que se incorpora a sociedades preexistentes: forma parte de un proceso de hibridación en el que ciencia, tecnología, cultura y política se transforman mutuamente. La velocidad con que los nuevos desarrollos pasan de la investigación a la apropiación social reduce, además, el tiempo disponible para comprender sus consecuencias, deliberar sobre ellas y construir respuestas institucionales. La posibilidad de anticipar los efectos de una innovación se vuelve así cada vez más limitada, porque las tecnologías adquieren usos y generan comportamientos que exceden con frecuencia los objetivos para los cuales fueron concebidas.

* Director de la Oficina de Organización de Estados Iberoamericanos para la Educación, la Ciencia y la Cultura (OEI) en Argentina.

Esta aceleración plantea un desafío particularmente relevante para la ética de la inteligencia artificial. No se trata solamente de establecer principios que orienten el diseño y el uso responsable de los sistemas algorítmicos, sino también de construir capacidades sociales e institucionales para deliberar sobre tecnologías cuyos efectos se despliegan en contextos complejos e inciertos. En este escenario, la gobernanza de la IA debe asumir la tarea de articular innovación y prudencia, velocidad y deliberación, competencia y equidad, intereses particulares y el bien común. Más profundamente, supone reconstruir un sentido de lo común en torno a aquello que queremos preservar de nuestra condición humana en un mundo en transformación. La dignidad, la inclusión, la autonomía y la solidaridad dejan entonces de ser solamente principios normativos para convertirse en criterios desde los cuales orientar las decisiones colectivas sobre el futuro tecnológico.

Esta transformación modifica fundamentalmente las relaciones de poder. La expansión de las tecnologías digitales ha dado lugar a nuevos actores con capacidad de incidencia global, sustentados, entre otros factores, en la acumulación de datos, la capacidad de intervenir sobre las pautas de comportamiento mediante algoritmos y el control de infraestructuras y servicios esenciales de conectividad. La inteligencia artificial profundiza esta tendencia y plantea interrogantes que exceden la relación entre usuarios y tecnologías: cómo se controlan los recursos necesarios para su desarrollo, cómo se definen las reglas de acceso y utilización, quién asume los potenciales costos, entre otros interrogantes.



Al mismo tiempo, la velocidad y la complejidad de estas transformaciones ponen a prueba las capacidades de las instituciones existentes. Muchas de estas organizaciones fueron concebidas para operar en contextos más estables, con tiempos de decisión más prolongados y con actores y responsabilidades relativamente delimitados. La aceleración tecnológica introduce, en cambio, situaciones en las que las decisiones deben tomarse con información incompleta y donde los efectos de una innovación pueden desplegarse antes de que existan marcos adecuados para comprenderlos o regularlos. La irrupción de la IA nos sitúa en un escenario donde se requieren instituciones capaces de aprender, adaptarse y coordinar actores diversos, sin renunciar a la legitimidad y a los principios que deben orientar su acción. El desafío no consiste solamente en regular una tecnología en permanente evolución, sino en desarrollar nuevas capacidades colectivas para gobernar la incertidumbre que ella contribuye a generar.

Sin un sistema institucional renovado y con un involucramiento plural y colaborativo, existe el riesgo de que surjan nuevas asimetrías, así como también un aumento considerable de la dependencia de los dispositivos, la soledad y el aislamiento de generaciones enteras, especialmente niños y adolescentes. Lo decisivo no es definir sencillamente cuánto puede crecer la IA, hasta dónde llegará el algoritmo en términos de cómputo y generación de información, sino bajo qué condiciones objetivas se hará realidad el formidable potencial que esta tecnología encierra dentro de sí.

Ante esta transición civilizatoria acelerada por las nuevas tecnologías, el futuro se presenta como una combinación de grandes oportunidades de mejora y la posibilidad de que estas mutaciones provoquen inestabilidades sistémicas en un grado todavía desconocido. Este horizonte nos obliga a repensar las relaciones entre instituciones, empresas y gobiernos para abordar la creciente complejidad con nuevos instrumentos que faciliten acuerdos donde la ética señale un camino para los adelantos que la IA promete y está en condiciones de cumplir.

# Labor, Productivity, and the Sabbatical Expectations of Artificial Intelligence

**Jordan Schwartz** *



The idea of Sabbath can be traced to Genesis when God declared rest after that very first week of worldly activity. What was the motivation for resting and how does it inform our expectations of work today? A lay interpretation of the origin generally emphasizes the word, "rest," but in Hebrew, the word *Shabat* is derived from the verb in the first phrase of Genesis 2:2: "God *finished* his work which he had done." That is, the blessings of rest come when we have *completed* our work, not simply when we need or want a break from it.

In modern liberal economics, the notion of quality of life—our capacity to enjoy the fruits of our labor, or the ability of labor to generate disposable time—are built on the principle of Productivity and are surprisingly consistent with the original meaning of the Sabbath. Economists ask, "What tools are at the disposal of industry and public policy so that our workers can become more productive—increase their output, and the earnings that come from that output—without having to devote more time to work?" In other words, "How can we *finish* our work more quickly so we can rest more?"

This fundamental question should impact everything from public investment plans to tax incentives, from training and education policies to labor laws. Workplace regulations that mandate minimum leave, maternity and paternity leave, holidays, sick leave, and even the construct of the work week and weekends are all forms of sabbatical, according to both the original and more generous interpretation of the word.

* Former Executive Vice President of the Inter-American Development Bank.

**Technology's role in productivity**

Before we set policies for work, we must understand what drives productivity. The traditional factors understood to drive the growth of economic output—labor and capital—do not explain *all* of the growth in our economies or firms from one year to the next, nor do they explain the years of contraction. Something is missing when we add labor and capital together. This x-factor is presumed by most economists to be the ability of an economy to absorb new technology, to derive more efficient ways of producing, conducting business, and offering services. When innovations allow us to do more without the need for more work hours or more capital investment, then economic growth needs to account for this ephemeral input. That input is "Productivity," and it generates output as economic growth on the one hand and room for workers to rest on the other.

To offer an example, if a research assistant can produce, say, three times as much information in a workday as that same worker could before internet access--and if an internet search engine costs less than two additional research assistants--then we need to include the contributions of that innovation when we look at economic output.

To estimate what the technology and innovation contributions are to economic growth, changes in Productivity are calculated as the *residual* of an economy's overall growth—what is left after we account for changes in the economy's labor and capital markets. Because we cannot measure it or see it at an aggregate level, we infer its existence from the wake that it leaves— much as we identified the planet Neptune in the nineteenth century based on disturbances in Uranus's orbit.



Technological revolutions thus bring us both expectations (higher growth) and aspirations (higher quality of living). Outside of academia, we experience productivity gains through observation, trying to apply what we have seen from *past* technology changes to the emerging technologies we are witnessing. We ask, "How will self-driving vehicles change the value of our commuting time?" "How will the use of battery storage, in combination with the sun and wind as sources of energy, impact our national accounts, or our economic efficiency, if they displace thermal power?" And more than ever, "Now that Artificial Intelligence can generate whole research papers, answer confused customers' most urgent questions, produce better forecasts, and spot both anomalies in deep outer space and hidden treasures under rainforest canopy, how will our Productivity change?" We look at what the Internet brought us and try to imagine what AI's marginal contribution will be to our own way of doing and building —and resting.

The expectation that comes from technological advances is that new levels of output and wealth, and, with them, disposable time, are generated by their benefits even when the other resources at hand remain fixed. This goes back to the invention of bronze, the wheel, the concept of "zero," antibiotics, fertilizers, washing machines. Did these inventions not give our societies greater productivity, the room to extend life expectancy, increase time with family, and more comfortable lives? Did they not give us the Sabbath we had earned?

The first corollary to the notion of asymmetric gain from new technology is the disruptive or "counter-productive" element of human behavior. Alfred Nobel discovered this contradiction when his invention of dynamite--which he thought would primarily revolutionize construction—was quickly deployed for political violence purposes. Hardly had prototype self-driving vehicles begun to emerge than guerrilla videos followed showing people standing over bridges, hacking into the cars, and propelling them to crash. On a larger scale, social media may be the early 21st Century's poster child for technologies that work against their own potential. One day, not long ago, we were using the new platforms to reconnect with old friends and crowdsource funds for new inventions, and seemingly the next day, a generation of mental health issues had emerged from cyber-bullying and a tsunami of self-obsessiveness.

**Artificial intelligence encounters disruptive human behavior**

Even if we are to assume that AI will bring more productivity gains than anti-social behavior, there remains a real risk that human nature will intercede to diminish the new technology's full potential. As the first hard-to-refute generated videos of political and social leaders speaking against their own interests begin to penetrate the social media landscape, we begin to imagine how AI can contribute further to the erosion of verifiable forms of information, and the diminution of institutions needed to propel economic activity and social productivity. That is, there is no reason to assume that we ourselves will not work against the gains that could have been fully inherent in the same technologies that are intended to produce growth and productivity premia. The hope remains that an arms race that pits AI-enabled virtuous policing against anti-social or vicious abuse of AI will protect the promise of machine learning, but it is hard to imagine that there will be no expenditure of time and resources in combatting the misuses of AI that are already emerging.



**AI vs. a free lunch**

A second constraint on the potential of AI to fulfill its productivity potential comes from the physical inputs required to deploy AI services. First, we see the massive amounts of capital required to build the chips that drive AI. We are already witnessing a concentration of equity market value capitalizations in mega-firms that are invested heavily in AI development and the associated chip manufacturing, a massive deployment of capital at historical levels which may or may not survive the next wave of technological change.

The other dislocation from physical inputs to AI can be found in the staggering volumes of energy required to run the data centers that fire AI's algorithms. The cost to utilities, consumers and corporations of having to increase investments and expenditure on energy brings its own disruption. By 2030 we expect AI to consume about one-fifth of all growth in energy demand. Depending upon a system's availability

of excess installed power generation, and the structure of the local power market, those costs may be hard to incorporate into current consumer fees. Indirectly but importantly, this increase in demand without a commensurate set of increases in supply of power generation, means competition among consumers and higher prices. Where power generation *can* be piled onto the systems quickly, it means more competition for scarce natural resources, and greater risk of stranded assets if the future forms of generating AI algorithms do not require as much power.

And yet…

Despite these two constraints—the behavioral, or anti-productivity risks, and the physical costs—, the original promise of AI can already be found across every field of natural and social science that depends on data analytics, modeling and forecasting, research, and problem solving. AI is already embedded in the workflow of a growing number of fields, and, in many cases, reshaping productivity.

**The distributional effects of AI**

The first labor market questions to arise are around the *distribution* of vulnerability to AI—for example, whether less skilled workers, and younger workers, are more likely to be displaced by AI than their wealthier and older counterparts. A large part of the answer is related to the redistribution of labor that began with the Industrial Revolution and accelerated in the late 19th and 20th centuries with large-scale automation of production. The massive growth in productivity in the agriculture sector since the start of the "Green Revolution" was driven by both the development of resilient seeds, fertilizers, pesticides and the automation of farming techniques. Mining and manufacturing have followed similar trends, and it appears that AI will contribute to this continued decline in labor inputs as a share of production, particularly in those countries with the educational and technological foundation to adopt rapidly changing modes of production.



Service sector jobs now provide most employment opportunities for unskilled and lower wage workers in higher income countries. The early data suggests that this will also be an area of vulnerability. For middle and upper middle income developing regions, such as Latin America and the Caribbean, the majority of the region's poor live in the wealthier countries of the region where service sectors —such as home care, logistics, warehousing and delivery, tourism, retail, and customer service back offices—employ a disproportionate share of low-skilled and semi-skilled workers, many informally employed.

Growth Theory and empirical evidence suggest that long-run productivity growth is driven not primarily by the accumulation of physical capital, but by the generation,

adoption, and spread of new ideas, skills, and organizational capabilities.[1] Viewed through this lens, large portions of the service sector may continue to struggle to become dynamic sources of growth. When service activities fail to integrate technological change, provide workforce training, or offer improved modes of organization, workers tend to remain confined to low-value tasks. This dynamic helps explain why poverty may endure even in relatively high-income economies: low-productivity services not only mirror existing inequalities, but can also reinforce them by constraining wage growth and limiting paths for social mobility.

**The countervailing promise of AI**

In the labor market, early analysis of real-world deployments show that generative AI systems capture the patterns of top performers and diffuse them to newer workers. In customer support, for example, randomized adoption of an AI assistant raised issues resolved per hour by about 15 percent on average, with the largest gains accruing to less-experienced agents—shortening the journey from apprenticeship to competence without expanding the workday.

When an AI tool was integrated into customer support interactions, average output increased by nearly 14 percent.[2] The largest improvements—around 35 percent—were concentrated among junior and lower-skilled agents, while top-performing and highly experienced workers experienced few or no gains. Similar field experiments in software development report over 25 percent more completed tasks among developers using AI code assistants, again with disproportionate benefits for junior staff. This creates a subtle paradox: younger and less experienced workers are often the quickest to integrate AI into their daily tasks and reap immediate productivity gains, yet they are also the most exposed to some of its side effects.



If the Sabbath follows the *finishing* of our work, then part of AI's promise should be inter-generational—the acceleration of productivity can be captured by the many—as both workers and as consumers, not only the few. This of course presumes that entire fields—such as customer support, radiology, architectural drafting, translation, or paralegal services, to name a few—are not displaced by AI so rapidly that we witness great dislocation costs at the sectoral and at the human level. At the very least, it presumes that the savings from reduced labor on those services accrue to other, yet unforeseen, work arenas. But these shifts are neither automatic nor instantaneous.[3]

1. See Romer (1994).
2. See Brynjolfsson *et al.* (2026), and Cui *et al.* (2025).
3. Jaffe *et al.* (2024) show that the influence of generative AI varies by role, function, and organization, and is contingent on adoption and utilization.

## The shape of AI's coming benefits?

The "J-curve" of complements suggests that intangible investments made today translate into measurable productivity gains later. To reap society-wide benefits, general-purpose technologies rarely deliver free lunches; they demand complementary investments—process redesign, data integrity and robustness, managerial routines, or new human capital—that national accounts often under-measure. The literature describes a *Productivity J-curve*, meaning an early dip in output while organizations absorb the costs of those hard-to-value complements, followed by a rise in productivity once those assets bear fruit.[4] Recent evidence in U.S. manufacturing finds short-run productivity losses, higher work-in-process, robot investment, and uneven impacts across firm types during early AI adoption.[5]

The above implies that the sabbatical of greater leisure is *deferred* while firms and workers both learn how to produce more efficiently. During the transition, the costs and benefits of productivity are distributed unevenly—across different workers, among firms within the same sector, and across firms in different sectors—unfolding with distinct temporal dynamics. One of those "uneven distributions" appears to be the way young workers are impacted versus older workers.

While the young coders and customer service support workers saw jumps in productivity vis workers less able to integrate AI into their work, early evidence suggests that as generative AI tools take over the drafting, summarizing, and templating once done by junior professionals, firms that have depended on those particular kinds of functions are beginning to hire fewer entry-level roles.[6] That said, they are shifting demand toward positions that require experience, judgment and domain stewardship. In software, postings for junior developers have fallen relative to senior roles since the public diffusion of GenAI, while field experiments show that quality and speed gains are "jagged," strongest where tasks align with AI's frontier and weaker where human context and verification are indispensable.[7]



## Preparing for AI

The way firms prepare for AI will have an outsized role in how the labor market in those fields develops. Labor policies can shape that preparation. Comprehensive, long-term labor policies that establish a balance between preserving career ladders (apprenticeships, co-ops, mentorship) and investing in skills that *complement* automation, so the benefits borne of higher Productivity do not foreclose the learning that makes future masters. There is no evidence yet that whole areas of economic activity can function without living

4. See Brynjolfsson *et al.* (2018).
5. McElheran *et al.* (2025) examine the prevalence and productivity dynamics of artificial intelligence in American manufacturing.
6. See Westby and Modestino (2025).
7. See Dell'Acqua *et al.* (2023).

memory—the uncodified knowledge, lived experience and relationships that drives our understanding of the markets in which we work—and, therefore, that AI can come to substitute for shared intergenerational knowledge among workers.

AI can boost productivity by improving decision-making, allowing full or partial automation of tasks and processes, while outcomes hinge on adoption depth and complements.[8] As with a cooking recipe, these complements are the key. They are what can either ruin a dish completely or elevate it into a work of art. The divine is also in the details.

The accelerated adoption of AI forces us to redouble our focus on preventing social fragmentation and youth unemployment. Intergenerational bridges between experienced workers and younger generations will protect the stability of our institutions. Robust pillars for these bridges, in turn, require some sharing of the dividends from the AI-generated sabbatical.

A balanced distribution among young people with differing access to educational opportunities. Forms of training need to recognize that even if entry level jobs seem most vulnerable, younger workers adapt faster and will benefit more from AI if given access to AI tools and training. Young workers already make up a disproportionately high share of generative AI use globally (e.g., 46% of ChatGPT messages are sent by users 18–25), indicating strong openness to AI augmented workflows. It is important to identify which junior tasks can become *AI-supported* rather than *AI-replaced* to ensure early-career staff remain owners of *judgment-based* components of work, supported by automation rather than displaced by it.



A balanced distribution between workers more and less exposed to automation. A humane response will center itself around preparation of the workforce, including investment in structured reskilling pipelines, especially where automation risk is highest. This suggests tailoring country-specific and sector-specific reskilling and upskilling programs to mitigate negative job effects and prepare workers for hybrid roles that combine human judgment with AI systems.

A balanced distribution between workers in industry and services. As AI adoption advances, productivity gains are unlikely to be evenly distributed across sectors, with early and larger effects concentrated on capital-intensive and technologically advanced activities. A policy-relevant response therefore requires deliberate efforts to extend productivity-enhancing technologies to service sectors, where a substantial share of low-wage and informal employment is concentrated. Facilitating the diffusion of AI tools, complementary skills, and improved organizational practices in services can help ensure that productivity growth does not widen existing sectoral gaps, but instead contributes to more balanced productivity gains across the economy.

8. A more detailed discussion is provided in IDB (2025).

**Public policy's role in addressing AI's pending challenges**

As AI functions expand, the importance of ethical, inclusive workforce planning grows alongside the technology. This means planning for responsible human-AI teaming, and alignment of human capital practices with AI-era realities. It also requires conducting structured workforce impact assessments to identify which junior and low-skilled functions face the highest automation exposure under the new technology; embed "future-skills hiring" criteria—valuing adaptability, reasoning, structured problem-solving—into entry-level recruitment, and ensure policies can mitigate, or prepare for, large-scale displacement. Neither behavioral nor physical constraints should pull us away from a goal that envisions the use of AI for the common good.

In both emerging and industrialized economies, it will become increasingly important to frame policy dialogue around the form in which governments, firms and workers interact with and leverage AI. AI's impacts will ultimately depend on this three-way dialogue—and on the intentionality that we bring to that dialogue. As with AI tools themselves, that coordination must be shaped, guided, and continually directed by human judgment.

**References**



Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. The Quarterly Journal of Economics, 140(2), 889-942.

Brynjolfsson, E., Rock, D., & Syverson, C. (2018). The productivity J-curve: How intangibles complement general purpose technologies. NBER Working Paper No. 25148.

Cui, K. Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S., & Salz, T. (2025). The effects of generative AI on high-skilled work: Evidence from three field experiments with software developers. Working paper, february.

Dell'Acqua, F., Rajendran, S., Krayer, L., Candelon, F., Lakhani, K. R., McFowland III, E., Mollick, E., Lifshitz-Assaf, H., & Kellogg, K. C. (2024). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. MIT Sloan School of Management Working Paper No. 24-013.

Inter-American Development Bank Group (2025). Artificial intelligence framework for the Inter-American Development Bank Group. IDB.

Jaffe, S., Parikh Shah, N., Butler, J., Farach, A., Cambon, A., Hecht, B., Schwarz, M., & Teevan, J. (2024). Generative AI in real-world workplaces. Microsoft Research Technical Report MSR-TR-2024-29.

McElheran, K., Yang, M.-J., Kroff, Z., & Brynjolfsson, E. (2025). The rise of industrial AI in America: Microfoundations of the productivity J-curve(s). U.S. Census Bureau Working Paper No. CES-25-27.

Romer, P. M. (1994). The Origins of Endogenous Growth. Journal of Economic Perspectives, 8(1), 3-22.

Westby, D., & Modestino, A. (2025). GenAI and software developer job vacancies 2. Manuscript submitted to Management Science (MS-0001-1922.65).

# ARTÍCULOS INVITADOS

Los artículos publicados fueron solicitados a sus autores por los expertos Ismael Gómez, director de estrategia digital global de la Organización de Estados Iberoamericanos (OEI) para la Educación, la Ciencia y la Cultura, y Diego Álvarez Newman, investigador del CONICET en el Instituto de Estudios Sociales en Contextos de Desigualdades de la Universidad Nacional de José Clemente Paz (UNPAZ), Argentina, y miembro del Grupo de Estudios sobre IA del Consejo Episcopal Latinoamericano (CELAM).

# Inteligencia artificial. La urgencia de una nueva ética en un contexto global de cambio

# Inteligência artificial. A urgência de uma nova ética em um contexto global de mudança

# *Artificial Intelligence. The Urgency of New Ethics in a Global Context of Change*

**Gabriela Agosto** *

El desarrollo acelerado de la inteligencia artificial (IA) constituye uno de los fenómenos más significativos y disruptivos del siglo XXI, tanto por su capacidad de transformar las estructuras que cohesionan la sociedad actual como por las consideraciones éticas que suscita su aplicación. En un contexto global en el que convergen simultáneamente múltiples crisis sociales, políticas, económicas y ambientales, la IA se presenta como promesa y amenaza a la vida como la conocemos. Su impacto no se limita a la esfera técnica, sino que afecta las dinámicas de poder, la organización y productividad del trabajo, la educación y socialización, la producción académica y la formulación de políticas públicas. Ante las inquietudes y los interrogantes que estos avances tecnológicos plantean, la reflexión ética emerge como condición imprescindible para que esta tecnología esté al servicio del ser humano, y evitar que se convierta en instrumento de exclusión o dominación. Este artículo propone un abordaje interdisciplinario desde las ciencias sociales, incorporando una reflexión metaacadémica: el uso de la IA como herramienta para analizarse a sí misma de manera de contribuir a los desafíos contemporáneos.

**Palabras clave**: inteligencia artificial; ética; algorética; colonialismo digital; soberanía algorítmica



* Decana de la Facultad de Ciencias Sociales, Educación y Comunicación de la Universidad del Salvador (USAL), Argentina. Doctora en ciencia política y sociología por la Universidad Complutense de Madrid (UCM), España. Magíster en administración pública (IUPOG-UCM). Licenciada en sociología por la Universidad de Buenos Aires (UBA), Argentina. Investigadora y docente universitaria, consultora en temas de desarrollo e integración regional en el ámbito nacional e internacional. Correo electrónico: gagosto@usal.edu.ar. ORCID: https://orcid.org/0009-0008-5842-6967. Para este artículo se contó con la colaboración de Lucio Aya Tenorio, licenciado en relaciones internacionales por la Universidad Católica Argentina (UCA) y académico de la Facultad de Ciencias Sociales, Educación y Comunicación de la Universidad del Salvador (USAL), cuya asesoría y reflexión crítica resultaron fundamentales para el desarrollo del trabajo. Correo electrónico: luciotenorio@usal.edu.ar. ORCID: https://orcid.org/0009-0007-0089-9908.

O desenvolvimento acelerado da inteligência artificial (IA) constitui um dos fenômenos mais significativos e disruptivos do século XXI, tanto por sua capacidade de transformar as estruturas que mantêm a sociedade contemporânea quanto pelas considerações éticas que suscita sua aplicação. Em um contexto global no qual múltiplas crises sociais, políticas, econômicas e ambientais convergem simultaneamente, a IA se apresenta como promessa e ameaça à vida tal como a conhecemos. Seu impacto não se limita à esfera técnica, mas também afeta as dinâmicas de poder, a organização e a produtividade do trabalho, a educação e a socialização, a produção acadêmica e a formulação de políticas públicas. Diante das preocupações e dos questionamentos que esses avanços tecnológicos colocam, a reflexão ética emerge como condição indispensável para que essa tecnologia esteja a serviço do ser humano e não se converta em instrumento de exclusão ou dominação. Este artigo propõe uma abordagem interdisciplinar a partir das ciências sociais, incorporando uma reflexão meta-acadêmica: o uso da IA como ferramenta para analisar a si mesma de modo a contribuir para os desafios contemporâneos.

**Palavras-chave:** inteligência artificial; ética; algorética; colonialismo digital; soberania algorítmica

*The rapid development of artificial intelligence (AI) constitutes one of the most significant and disruptive phenomena of the 21st century, both for its capacity to transform social structures and for the ethical considerations that its application raises. In a global context where multiple social, political, economic, and environmental crises converge simultaneously, AI appears as both a promise and a threat to life as we know it. Its impact is not only limited to the technical sphere but also affects power dynamics, the organization and productivity of labor, education and socialization, academic production, and the formulation of public policies. Faced with the concerns and questions raised by these technological advances, ethical reflection emerges as an essential condition to ensure that this technology serves humanity and does not become an instrument of exclusion or domination. This article proposes an interdisciplinary approach from the social sciences, incorporating a meta-academic reflection: the use of AI as a tool to analyze itself to contribute to contemporary challenges.*

***Keywords:*** *artificial intelligence; ethics; algorethics; digital colonialism; algorithmic sovereignty*

## Introducción

La tercera década del siglo XXI fue testigo de una aceleración sin precedentes en los procesos de innovación tecnológica, cuyo exponente más paradigmático es la emergencia de la inteligencia artificial (IA). Su desarrollo, expansión y apropiación social se producen en un tiempo histórico caracterizado por la incertidumbre, la fragilidad de las instituciones y la velocidad del cambio. Si Bauman (2000) habló de modernidad líquida, autores como Riorda (2021) han propuesto recientemente que el paradigma sociohistórico actual es más bien gaseoso, volátil, impredecible e intangible: las estructuras y referentes tradicionales se desvanecen antes de que sea posible aferrarse a ellos o comprender su impacto e influencia.

En paralelo, la humanidad enfrenta lo que Morin (2020) y Toose (2021) describen como una policrisis: una multiplicidad de crisis simultáneas e interconectadas (climática-ambiental, económica, social, política y cultural) que no pueden ser comprendidas ni abordadas de manera aislada. La IA irrumpe en un escenario de vulnerabilidades acumuladas, en el que el desajuste entre los problemas sociales no resueltos, la crisis de la democracia liberal, los procesos bélicos y el vertiginoso avance tecnológico se convierte en un signo de nuestro tiempo.

En este contexto, el debate sobre la ética de la IA deja de ser una cuestión secundaria o meramente técnica. Su dualidad, con sus enormes potencialidades y sus riesgos igualmente significativos, obliga a reflexionar en sus aplicaciones inmediatas y en sentido último de su integración en la vida humana. Las ciencias sociales, al analizar los impactos de la IA en las interacciones, la producción académica y las políticas públicas, se enfrentan al desafío de elaborar marcos teóricos, normativos y regulatorios capaces de acompañar una transformación de alcance global. Este artículo parte de la premisa que la ética de la IA no puede pensarse solo desde marcos nacionales ni desde perspectivas estrictamente tecnológicas. Se requiere un enfoque interdisciplinario que reconozca a la humanidad como un todo, más allá de las fronteras y los particularismos.



Finalmente, este texto incorpora como epílogo una reflexión hecha por la IA, con el objetivo de explorar sus alcances y limitaciones en su autoconocimiento, haciendo de la escritura misma un espacio de reflexión crítica. De este modo, la investigación se despliega no solo sobre la IA, sino también con ella, invitando a considerar las implicancias éticas, sociales y filosóficas de un futuro en el que la inteligencia humana y la artificial coexisten.

De esta manera, se propone, en primer lugar, ofrecer un panorama general que dé cuenta de la complejidad y la dificultad de abarcar los impactos de la IA en todas las dimensiones socioculturales posibles. En segundo lugar, busca proponer un espacio de reflexión sobre la construcción colectiva de una nueva ética que oriente su desarrollo, como una tarea urgente e impostergable. En este sentido, se sostiene que la IA y otros avances tecnológicos recientes provocaron una ruptura del contrato social

tal como lo entendimos desde una visión rousseauniana. De allí surge la necesidad de elaborar un nuevo contrato, capaz de responder a los desafíos de esta transición. La pregunta que subyace es si la IA constituye un nuevo tipo de Leviatán, una entidad tan poderosa que, lejos de limitarse a servir al ser humano, se erige como mediadora y condicionante de sus vínculos, sus instituciones y su futuro común.

Para los fines de este trabajo, nos referiremos a la IA en su generalidad, sin realizar distinciones exhaustivas entre los múltiples tipos, modelos, servicios, capacidades y características específicas de los diversos productos ofrecidos por las corporaciones que lideran el sector como OpenAI (ChatGPT), Google DeepMind (Gemini), Anthropic (Claude), Meta (LLaMA), Microsoft, IBM o Amazon Web Services). Tampoco haremos una separación sistemática entre la IA generativa y la IA predictiva.[1] Lo que interesa no son tanto las diferencias técnicas entre ellas, sino lo que comparten desde un punto de vista epistemológico: la capacidad de procesar información, producir representaciones y generar efectos sociales que interpelan nuestra comprensión de lo humano. Esta mirada permite situar el debate en el nivel de las transformaciones estructurales. No obstante, conviene precisar una distinción conceptual: no se trata de la proliferación de "modelos específicos" en sentido estricto, sino de modelos generalistas contextualizados. Los LLM (*large language models*) como GPT, Claude o Gemini constituyen infraestructuras de propósito general que, mediante ajustes de contexto, interfaces o *fine-tuning*, se orientan a usos concretos (desde el ámbito jurídico hasta la salud, la educación o la gestión administrativa). Lo que varía, entonces, no es el modelo de base, sino el ecosistema de datos, aplicaciones y normativas en el que se inserta. Esta aclaración es fundamental para evitar lecturas que sobrestimen la diversidad de modelos y subestimen la centralidad de unas pocas arquitecturas dominantes en el mercado.



## 1. Revisión de literatura y estado del arte

La historia de la inteligencia artificial se remonta a mediados del siglo XX, con hitos importantes en la década de 1960. Sin embargo, el punto de inflexión se dio a fines de 2022, cuando OpenAI lanzó ChatGPT. A diferencia de los desarrollos previos, cuya circulación permanecía restringida a ámbitos especializados, ChatGPT democratizó el acceso a LLM, integrando sus capacidades en la vida cotidiana de millones de personas en cuestión de meses. Este salto cualitativo no radica únicamente en la potencia técnica del modelo, sino en su apropiación social masiva, que lo convierte en un fenómeno cultural y político, además de tecnológico.

Esta irrupción, así como su desarrollo previo, dio lugar a un *corpus* creciente de reflexiones que trascienden lo meramente técnico para situarse en el plano social,

[1] La primera puede definirse como aquella capaz de producir nuevos contenidos (textos, imágenes, sonidos, código) a partir de patrones aprendidos en grandes volúmenes de datos; la segunda, en cambio, se centra en inferir probabilidades y anticipar necesidades o comportamientos futuros a partir de datos históricos.

ético y político. La literatura especializada coincide en que el impacto de la IA es transversal, de manera que afecta por igual a la economía, la política, la cultura, la vida cotidiana y la esfera íntima de las personas. Su carácter disruptivo motiva la aparición de una diversidad de enfoques provenientes de la filosofía, la sociología, la ética, las ciencias políticas y la teología, que buscan comprender tanto sus potencialidades como sus riesgos.

En la literatura actual sobre ética e IA convergen diversas líneas de análisis, para este documento se seleccionaron tres, entendiendo que conceptualizan de manera integral la problemática. En primer lugar, la evidencia empírica sobre daños sociales causados por la IA (sesgos algorítmicos, discriminación automatizada, vigilancia masiva, daño ambiental), documentada en estudios de autores como Crawford (2021) y Zuboff (2019), que muestran cómo estas tecnologías pueden profundizar desigualdades preexistentes. En segundo lugar, los marcos normativos en desarrollo que buscan trasladar los valores de transparencia, responsabilidad e inclusión a criterios operativos y exigibles, como la *Rome Call for AI Ethics* (2020), la *AI Act* (2024) de la Unión Europea (UE) y los principios de la UNESCO (2021) y la Organización para la Cooperación y el Desarrollo Económico (OCDE, 2019, actualizados en 2024). En tercer lugar, la investigación emergente en justicia algorítmica, encabezada por académicos quienes desarrollaron metodologías para medir, evidenciar y corregir las desigualdades reproducidas por sistemas automatizados. En este marco, el concepto de "algorética" (Francisco, 2020) se presenta como un lenguaje común que traduce preocupaciones éticas universales vinculadas a la dignidad humana, la equidad y el respeto a los derechos fundamentales en principios aplicables al diseño y la regulación.



Los sistemas de IA generativa y modelos fundacionales desplazaron la discusión desde la promesa tecnológica hacia la gobernanza de riesgos sociales y civilizatorios. En términos normativos, la *Recomendación sobre la ética de la inteligencia artificial* de la UNESCO (2021) estableció un estándar global de referencia sobre derechos humanos, supervisión humana, transparencia, y rendición de cuentas, hoy utilizado por Estados y universidades como guía de políticas institucionales; los Principios de la OCDE (2019, actualizados en 2024) convergen en valores de derechos humanos, transparencia, robustez y responsabilidad y en recomendaciones de política pública, incluyendo cooperación internacional y preparación para la transición laboral. El 2024 marcó un punto de inflexión con la adopción del *AI Act* en la UE, que inauguró un esquema de obligaciones graduadas orientado a una IA "humano-céntrica y confiable" y a la protección de derechos fundamentales y de la democracia (Reglamento 2024/1689, considerandos 1-3).

La discusión sobre soberanía algorítmica condensa preocupaciones de países y regiones por asegurar capacidad de decisión, infraestructura y datos frente a las crecientes dependencias tecnológicas. En Europa, el *AI Act* se acompaña de un ecosistema de implementación (AI Office, *sandboxes*, códigos de práctica) que busca compatibilizar innovación con protección de derechos. En América Latina, se consolida una agenda de soberanía digital con debates sobre nubes públicas, estándares

abiertos, alfabetización de IA y compras públicas estratégicas, conectando con la historiografía regional e iniciativas de observatorios y redes académicas que mapean riesgos y oportunidades. Sin embargo, no se hicieron grandes progresos respecto de la regulación de esta tecnología.

En el campo de la filosofía y la ética, Floridi (2013, 2019) planteó la necesidad de una *ética de la información* y de un marco normativo que acompañe el desarrollo de las tecnologías digitales, en particular la IA, al sostener que el problema central no es solo lo que las máquinas pueden hacer, sino cómo sus capacidades reconfiguran las condiciones de posibilidad de la vida humana y social. En la misma línea, Metzinger (2021) advierte sobre los riesgos de avanzar en sistemas de IA cada vez más autónomos sin una base ética sólida, alertando sobre la tentación de delegar decisiones morales en algoritmos que carecen de experiencia subjetiva y emotiva.

Desde la sociología y la ciencia política, Zuboff (2019) describió el fenómeno del "capitalismo de la vigilancia", mostrando cómo las plataformas digitales y los algoritmos transforman los datos personales en la materia prima de un nuevo régimen económico basado en la predicción y manipulación del comportamiento. Srnicek (2016), por su parte, analiza el ascenso del "capitalismo de plataformas", donde la concentración de datos y poder en pocas corporaciones genera nuevas asimetrías de poder global. Estas lecturas son fundamentales para comprender el trasfondo económico y político en el que se desarrolla la IA.



Crawford (2021) sostuvo que la IA no es simplemente un conjunto de sistemas técnicos, sino una infraestructura sociotécnica que refleja y amplifica las relaciones de poder existentes. Su análisis evidencia cómo los sesgos algorítmicos y los sistemas de clasificación automatizados perpetúan desigualdades raciales, de género y socioeconómicas. Eubanks (2018) llega a conclusiones similares al mostrar cómo la automatización en programas sociales en Estados Unidos llevó a nuevas formas de exclusión y marginación, lo que plantea interrogantes sobre el uso de la IA en políticas públicas.

La dimensión teológica y humanista también aporta al debate contemporáneo. El papa Francisco (2020), en su *Llamada de Roma*, introdujo el concepto de "algorética", que articula seis principios fundamentales (transparencia, inclusión, responsabilidad, imparcialidad, fiabilidad, seguridad y privacidad) para garantizar que la IA permanezca al servicio del ser humano. Ciucci (2023), desde la Pontifical Academy for Life, insistió en que la IA plantea un desafío civilizacional que obliga a pensar en términos de dignidad humana universal, más allá de particularismos nacionales o locales. Esta perspectiva coincide con las advertencias de Benanti (2022), teólogo y miembro del Vaticano en debates sobre IA, quien señala que el verdadero riesgo es la subordinación de lo humano a la lógica técnica, despojando a la sociedad de su capacidad de autodeterminación y reflexión moral.

Adicionalmente, ante la capacidad transformadora de la IA tanto en los ámbitos económicos y políticos como respecto a la concepción misma de la libertad y la agencia

humana, Harari (2018) advierte que los algoritmos podrían conocernos mejor que nosotros mismos y condicionar nuestras decisiones más íntimas, lo que cuestiona la noción misma de libre albedrío. Por otro lado, Cristianini (2023) propone entender la IA no como una inteligencia autónoma, sino como una extensión de los sistemas humanos de información, lo que desplaza el debate hacia la responsabilidad humana en su diseño y aplicación. Autores como Noble (2018) evidenciaron cómo la IA y los motores de búsqueda no son neutrales, sino que reproducen y amplifican discriminaciones, abriendo una línea de investigación crucial en torno a la justicia algorítmica.

El debate contemporáneo sobre sesgos algorítmicos se centra en los daños distribuidos y las arquitecturas de injusticia digital. Si bien el trabajo de Buolamwini y Gebru (2018) marcó un punto de referencia al demostrar el impacto diferenciado de los sistemas de reconocimiento facial según género y color de piel, la literatura actual destaca que estos problemas no son anomalías corregibles únicamente con "más datos" o "mejores modelos", sino expresiones de estructuras sociales preexistentes codificadas en sistemas técnicos e informáticos. Asimismo, Crawford y Paglen (2019) demostraron cómo las bases de datos de entrenamiento visual se construyen sobre categorías culturalmente situadas que naturalizan prejuicios históricos. En su investigación, Noble (2018) observó cómo los motores de búsqueda reproducen estereotipos raciales y de género al indexar y jerarquizar información, ilustrando que el sesgo reside en las bases de datos y en los modelos de negocio que los sustentan. Benjamin (2019) acuñó el concepto de "Nuevo Código Jim" para señalar cómo la innovación digital, presentada como imparcial, puede reforzar prácticas de segregación y exclusión racial. Desde otra perspectiva, Birhane (2021) propone un giro hacia marcos relacionales que abandonen la obsesión por métricas aisladas y se centren en el contexto, las relaciones de poder y las dimensiones éticas del cuidado. Esto implica reconocer que los sesgos no pueden eliminarse en abstracto, sino que deben abordarse como problemas situados en realidades sociales complejas.



A estos enfoques se suman contribuciones como la de Selbst *et al.* (2019), quienes advierten sobre la "falacia de abstracción" en la ética de la IA: la tendencia a analizar dilemas en términos excesivamente formales, descontextualizados de dinámicas sociales. De esta manera, la justicia algorítmica no puede reducirse a ajustes técnicos, sino que exige un replanteamiento del modo en que producimos, clasificamos y utilizamos datos en la sociedad global contemporánea.

Resulta evidente que existe desde la academia un consenso parcial, según el cual la IA, por su carácter disruptivo y su complejidad técnica y filosófica, no puede ser analizada de manera aislada, sino como un fenómeno civilizacional que afecta los cimientos mismos de la sociedad global.

## 2. Marco teórico

La noción de inteligencia ha sido objeto de múltiples aproximaciones a lo largo de la historia de la psicología, la filosofía y las ciencias cognitivas. En la tradición psicométrica,

Spearman (1904) introdujo el concepto de "factor g", entendido como una capacidad general de razonamiento subyacente a la ejecución de tareas cognitivas diversas. Este planteo, ampliado en los modelos jerárquicos como el de Cattell (1941), Horn (1968) y Carroll (1993), permitió construir herramientas de evaluación comparables, pero redujo la inteligencia a lo cuantificable. En contraste, enfoques cualitativos como las inteligencias múltiples de Gardner (1983) o la teoría triárquica de Robert Sternberg (1985, 1988) reivindicaron la pluralidad de habilidades, incluyendo dimensiones creativas y prácticas que escapan a los *tests* tradicionales como la consideración del coeficiente intelectual.

Desde la psicología del desarrollo, Piaget (1952) concibió la inteligencia como un proceso de equilibración adaptativa, mientras que Vygotsky (1978) subrayó la dimensión social y cultural de las funciones cognitivas superiores. Estos aportes abrieron el camino a la visión contemporánea de la inteligencia como un fenómeno situado y mediado por la interacción con otros y con las herramientas disponibles y potenciales.

En el terreno filosófico y cognitivo, Turing (1950) propuso su célebre prueba de imitación como criterio funcional para atribuir inteligencia a una máquina, desplazando la pregunta ontológica hacia la verificabilidad conductual. Su propuesta fue criticada por Searle (1980), quien en su experimento mental de la "habitación china" argumentó que manipular símbolos no implica comprensión semántica. Otros autores como Dennett (1987) sostienen que basta con adoptar la postura intencional para atribuir creencias y deseos a sistemas complejos, mientras que enfoques enactivistas como los de Varela, Thompson y Rosch (1991) y Noë (2004) entienden la inteligencia como una práctica encarnada y emergente de la interacción organismo-entorno.



En la ingeniería, la inteligencia se define a menudo como la capacidad de un agente para maximizar objetivos en contextos inciertos. Russell y Norvig (1995, 2020) la conceptualizan en términos de agentes racionales, mientras que Hutter (2005, 2007) y Legg (2007) propusieron la "inteligencia universal" como capacidad general de alcanzar metas en una amplia variedad de entornos. Estas aproximaciones ofrecen operacionalización formal, pero corren el riesgo de desatender dimensiones sociales, éticas y culturales.

Respecto del carácter "artificial" y los debates sobre lo que implica la inteligencia (es decir, si realmente se trata de una "inteligencia" artificial o es más bien un algoritmo predictivo sin capacidad real de creación), se advierte que la capacidad de "parecer humano" de sistemas de lenguaje como GPT o los modelos de creación de imágenes y videos no equivale a comprensión ni a fiabilidad epistémica. Tal como advierte Mitchell (2019), los modelos de IA actuales no "entienden" el mundo en el sentido humano del término, sino que manipulan correlaciones estadísticas derivadas de grandes volúmenes de datos.

Esta distinción es crucial, porque el riesgo de antropomorfizar la tecnología lleva a sobrestimar sus capacidades y a invisibilizar tanto sus limitaciones estructurales

como sus consecuencias sociales. En esta línea, Bender, Gebru, McMillan-Major y Shmitchell (2021) sostienen que las aplicaciones que buscan imitar de manera creíble a los humanos tienen un riesgo de daños extremos, como la reproducción de sesgos sociales, la generación de desinformación y los costos ambientales derivados del entrenamiento de modelos de gran escala. Su crítica apunta a que la fluidez lingüística no debe confundirse con comprensión semántica, apropiación conceptual o autoridad epistémica.

A partir de estas advertencias, en los últimos años se consolidó un giro metodológico hacia la gobernanza de los sistemas de IA. Mitchell *et al.* (2019) propusieron las "cartas modelo", fichas estandarizadas para documentar los supuestos, usos previstos y limitaciones. De forma complementaria, Gebru *et al.* (2018) desarrollaron las *datasheets for datasets*, un mecanismo para transparentar el origen, la composición y la calidad de los datos empleados en el entrenamiento. Estas herramientas no solo buscan aumentar la trazabilidad y la responsabilidad, sino que fueron incorporadas en marcos institucionales de evaluación de impacto algorítmico, especialmente en Europa y Estados Unidos.

Para Simon (1969), lo artificial se refiere a los artefactos orientados a fines que median entre un estado actual y uno deseado bajo ciertas restricciones. Simondon (1958), en cambio, propuso pensar los objetos técnicos como entes en proceso de individuación, con una historia interna y una evolución que los conecta con sus entornos humanos y no humanos. En la misma línea, la posfenomenología de Ihde (1990) y Verbeek (2011) sostiene que las tecnologías median nuestra experiencia y constituyen parte de la condición humana, borrando la frontera rígida entre lo natural y lo artificial.



Desde la sociología de la ciencia, Latour (1987) subrayó que los artefactos deben ser entendidos como actores que participan en redes híbridas de agencia. A su vez, la cibernética de Wiener (1948) y la teoría de la información de Shannon (1948) vincularon la artificialidad con la capacidad de diseñar sistemas de control y retroalimentación.

En definitiva, resulta posible extraer tres rasgos clave de la artificialidad: teleológicamente, se encuentra orientada a fines explícitos, posee una normatividad incorporada desde la cual materializa valores y decisiones humanas en diseños técnicos, y tiene la capacidad de transformar y ser transformada por prácticas sociales, instituciones y culturas. A partir de estos debates, proponemos definir la IA como una herramienta con capacidad para predecir y reaccionar de manera adaptativa en contextos inciertos, materializando iniciativas humanas en arquitecturas técnico-informacionales que impactan en todas las dimensiones de la vida y su medio consolidándose también como servicio transable con características extractivas. Esta definición combina dimensiones cognitivas (adaptación, inferencia, acción), técnicas (diseño, finalidad, información), económicas (cambio de paradigma del mundo del trabajo, la comercialización, la competencia) y socioambientales (valores, instituciones, coevolución, extractivismo).

### 3. Ética, civilización y policrisis en el antropoceno digital

La emergencia de la IA debe analizarse en el marco de un cambio civilizatorio más amplio. Ubicar la IA en el período del "antropoceno" implica reconocer que sus riesgos y promesas están entrelazados con las crisis planetarias actuales: el cambio climático, la pérdida de biodiversidad, la pobreza extrema, la precarización del trabajo y la fragmentación política. Vivimos en una época donde la distinción entre lo natural y lo social se desdibujó: las infraestructuras digitales son al mismo tiempo infraestructuras ecológicas y políticas. En este escenario, pensar una ética de la IA equivale a repensar el contrato social y ecológico de la humanidad. Este horizonte sitúa a la humanidad como fuerza geológica capaz de transformar radicalmente el planeta, pero también como especie que enfrenta sus propios límites.

El plano de la ética pública se configura como un cruce de normativas internacionales, propuestas de gobernanza y tradiciones filosófico-humanistas que buscan frenar la deriva tecnocrática. De allí que organizaciones internacionales como Naciones Unidas (ONU) proponen marcos normativos universales para la IA, que buscan articular principios comunes en torno a la justicia, la sostenibilidad y los derechos humanos. Los principios internacionales mencionados constituyen marcos de referencia que enfatizan derechos humanos, transparencia y rendición de cuentas. El *AI Act* de la UE introduce, además, un enfoque regulatorio *ex ante*, estableciendo categorías de riesgo y obligaciones específicas para los proveedores de sistemas.



Sin embargo, más allá de la regulación positiva, diversos autores advierten que la discusión debe anclarse en horizontes civilizatorios más amplios, superando las fronteras físicas, mentales, espirituales y cosmovisionales. Floridi y Cowls (2019) formularon los "cinco principios de la IA responsable" (beneficencia, no maleficencia, autonomía, justicia y explicabilidad) como una "gramática ética transversal". Desde el Sur Global, Arun (2020) cuestiona la suficiencia de los marcos eurocéntricos, subrayando la necesidad de incorporar nociones de justicia distributiva y epistemologías plurales. En este contexto, la algorética (Francisco, 2020) ocupa un lugar destacado al ofrecer una síntesis práctica y humanista que traduce la preocupación por la dignidad humana en criterios operativos para situar al ser humano en el centro del proceso tecnológico. La perspectiva humanista-cristiana, expresada por el papa Francisco, considera a la IA como una herramienta cuyo sentido depende de la orientación política y ética que le demos. En su mensaje en la cumbre del G7 en 2024, Francisco insistió en que "los beneficios o daños que traerá dependen de su uso", llamando a una "responsabilidad política" que vincule innovación tecnológica con bien común y no solo con beneficio económico.

Este marco abre la puerta a un debate más amplio sobre sostenibilidad ética: ¿cómo garantizar que el diseño y uso de la IA no solo no vulneren la dignidad humana, sino que promuevan activamente condiciones de justicia social, cuidado mutuo y equidad intergeneracional? La consolidación de una "ética de la IA", en consecuencia, se juega tanto en la creación de normas e instituciones como en la construcción de

culturas políticas y sociales que prioricen la centralidad de la persona frente a la lógica de la acumulación y el control.

Archibugi (2022) introduce el concepto de "soberanía tecnológica", que en el caso de la IA se traduce en soberanía algorítmica; es decir, la capacidad de los Estados y las sociedades para decidir sobre el diseño, la implementación y la gobernanza de los sistemas de IA. La concentración del desarrollo de IA en un pequeño número de corporaciones y países (principalmente en Estados Unidos y China) genera una dependencia global que amenaza con convertir a otras regiones en meros consumidores tecnológicos, sin voz ni control sobre las decisiones que configuran estos sistemas (Bender *et al.*, 2021). La discusión apunta, por tanto, a la necesidad de consolidar un marco ético que trascienda lo instrumental. Desde la algorética, los principios universales de transparencia, inclusión, responsabilidad, imparcialidad, fiabilidad y seguridad se proponen como pilares de una civilización digital al servicio de la dignidad humana.

Desde un enfoque laico, pensadores como Zuboff (2019) y Couldry y Mejías (2019) concluyen que la ética de la IA no puede reducirse a un código de buenas prácticas empresariales, sino que requiere un pacto civilizacional donde la tecnología sea subordinada al bien común. En palabras de Zuboff: "la elección es entre una sociedad en la que los individuos todavía pueden decidir sobre su futuro, o una donde las decisiones estén predestinadas por sistemas de cálculo invisibles" (2019, p. 514). Parafraseando a Helbing (2022), el futuro de la IA es el futuro de nuestra civilización; regular es regularnos a nosotros mismos.



Asimismo, los desarrollos recientes de los modelos de aprendizaje profundo y generativo a gran escala demostraron tener un impacto ambiental significativo debido a sus elevados requerimientos computacionales y energéticos. El entrenamiento de modelos como GPT-3 o GPT-4 demanda millones de horas de procesamiento en centros de datos que consumen cantidades masivas de electricidad y agua para refrigeración. Según Strubell, Ganesh y McCallum (2019), el entrenamiento de un único modelo de procesamiento de lenguaje natural puede emitir hasta 284 toneladas de $CO_2$, cifra comparable a las emisiones anuales de cinco automóviles promedio en Estados Unidos. Estudios posteriores como los de Patterson *et al.* (2021), desde Google Research, advierten que el costo energético de la IA aumenta exponencialmente con el tamaño de los modelos y se concentra en infraestructuras tecnológicas ubicadas en países desarrollados, lo que profundiza desigualdades ambientales y de acceso, así como la deuda climática con los países en vías de desarrollo. Además, Bender *et al.* (2021) señalan que el ciclo de vida de estas tecnologías, desde la extracción de minerales raros hasta el consumo de agua en *data centers*, contribuye de manera indirecta a la aceleración del cambio climático, ampliando la huella ecológica digital.

Este vínculo entre IA y degradación ambiental coloca en el centro la discusión ética sobre la sostenibilidad y la justicia intergeneracional. El entrenamiento y despliegue de IA contribuyen de manera verificable al incremento de las emisiones de gases de

efecto invernadero y al agotamiento de recursos hídricos y minerales. El principio de sostenibilidad, defendido por autores como Daly (1996), y más recientemente por la ONU en su declaración "Transformar nuestro mundo: la Agenda 2030 para el desarrollo sostenible", exige reconsiderar las trayectorias de desarrollo tecnológico y de su uso cotidiano bajo el imperativo de garantizar que las generaciones futuras puedan acceder a recursos y condiciones ecológicas mínimas para una vida digna. En este sentido, la aceleración del cambio climático, vinculada al uso masivo de energías fósiles en la expansión de *data centers* y redes de comunicación, convierte a la IA en un factor que no sólo refleja, sino que también amplifica la dimensión ecológica de la policrisis.

## 4. Economía política de la IA: extracción, plataformas y colonialidad de datos

La investigación interdisciplinaria contemporánea problematiza la IA como una infraestructura extractiva de alcance global. En esta línea, Crawford (2021) y Pasquinelli y Joler (2020) sostienen que se asienta sobre una trama material y laboral que suele quedar invisibilizada en los discursos celebratorios de los avances técnicos. La explotación de minerales estratégicos, el consumo energético intensivo y el trabajo precario en etiquetado de datos y moderación de contenidos constituyen los cimientos de sistemas presentados como "inmateriales" o "virtuales". El desarrollo y la subsistencia de la IA, en este sentido, no puede comprenderse sin atender a las cadenas de suministro y a los regímenes de extracción que la hacen posible.



En un plano más amplio, Zuboff (2019) conceptualiza este proceso dentro del marco del "capitalismo de vigilancia", entendido como un régimen económico que convierte la experiencia humana en materia prima para la predicción y el control conductual. Según la autora, las plataformas digitales operan como dispositivos de captura sistemática de información que, a su vez, alimenta sistemas algorítmicos destinados a modelar, anticipar y condicionar la conducta de los individuos. Esta dinámica erosiona la autonomía personal y transforma la deliberación pública en un espacio cada vez más mediado por intereses corporativos y lógicas de rentabilidad.

Esta perspectiva dialoga con la noción de "colonialismo de datos" (Couldry & Mejías, 2019), según la cual la apropiación y mercantilización transnacional de datos reproduce las mismas lógicas de despojo y subordinación que caracterizan al colonialismo histórico. Las poblaciones son fuente de recursos y también objeto de experimentación y control, incrementando la asimetría en la distribución de poder entre el Norte Global y el Sur Global. La IA, bajo esta óptica, se convierte en un vector de nuevas dependencias tecnológicas que limitan la soberanía de los Estados y profundizan desigualdades estructurales.

En paralelo, la trayectoria de la IA en su fase de apertura al público repite, con matices, la genealogía del propio internet. En sus primeras etapas, los modelos generativos fueron ofrecidos en modalidad gratuita, bajo la lógica de acceso abierto

y con el objetivo implícito de acelerar su perfeccionamiento a través de la interacción masiva con usuarios. Este proceso de entrenamiento "socializado" permitió que los sistemas aprendieran de los insumos lingüísticos, conductuales y contextuales aportados por millones de personas, convirtiendo a los usuarios no en consumidores pasivos, sino en coproductores involuntarios del desarrollo tecnológico. Sin embargo, una vez que los modelos alcanzaron un grado de maduración y estabilidad comercial, comenzó a imponerse la lógica del capitalismo de plataformas (Srnicek, 2016): la gratuidad inicial se restringió y emergieron versiones *premium*, modalidades de suscripción y *paywalls* que segmentan el acceso según capacidad de pago. Esta transición plantea una contradicción entre el acceso abierto y la democratización del conocimiento, y la monetización intensiva que amenaza con producir un sistema de exclusión digital, en el que solo quienes pueden pagar acceden a las funcionalidades más potentes y avanzadas.

Este fenómeno se entrelaza con una tendencia más amplia de la economía digital contemporánea: la plataformización de la vida cotidiana. Así como los *smartphones* y las aplicaciones fragmentaron las interacciones en múltiples microtransacciones diarias (desde la movilidad hasta la educación y la salud), la IA tiende a integrarse en todos los ámbitos de la vida pública y privada, convirtiéndose en un requisito de funcionamiento básico. El paralelismo con el ecosistema de *apps* tras la pandemia es evidente: para realizar trámites, consumir cultura o incluso mantener lazos sociales, el individuo depende cada vez más de infraestructuras digitales privadas. En este proceso, los grupos más vulnerables sin una política activa formativa (adultos mayores, personas con discapacidades, sectores sin alfabetización digital) quedan en situación de marginalidad estructural. Van Dijck, Poell y De Waal (2018) conceptualizan la "sociedad de plataformas" como un entramado sociotécnico en el cual se constituye en la infraestructura dominante de intermediación en sectores clave como la educación, la salud y los medios de comunicación. Estos procesos desplazan progresivamente funciones estatales y sociales hacia actores privados, generando tensiones respecto a la gobernanza democrática.



Entonces, la socialización inicial de la IA a través de interacciones abiertas plantea interrogantes sobre el futuro de la relación entre tecnología y sociedad. ¿Persistirá un régimen híbrido, con versiones gratuitas limitadas junto a servicios *premium*, o avanzaremos hacia un modelo de exclusividad creciente, donde las herramientas de mayor valor se reserven para élites corporativas, gubernamentales y económicas? La respuesta a esta pregunta determinará, en gran medida, si la IA reforzará las desigualdades sociales existentes o si, por el contrario, se convertirá en una herramienta democratizadora.

Por otro lado, el empleo y acceso masivo a la IA puede favorecer la utilización para la manipulación política y la expansión de esquemas de desinformación. El poder de la IA puede ser instrumentalizado tanto por actores individuales como por corporaciones de todo tipo. La capacidad de generar contenidos falsos hiperrealistas (*deepfakes*) plantea riesgos para la calidad democrática y los procesos electorales (Helbing,

2022). Los casos de Cambridge Analytica y la injerencia en procesos electorales como el Brexit o las elecciones estadounidenses de 2016 muestran cómo la manipulación de datos y flujos de información puede alterar la opinión pública. Con la generación masiva de imágenes, audios y videos falsos, la sociedad tendrá mayores desafíos para discernir lo real de lo ficticio.

También se estima que la evolución de la IA suscita una considerable pérdida de fuentes y puestos de trabajo, así como un impulso a la precarización laboral. La OCDE (2024) plantea que cerca del 27% de los empleos en países miembros enfrenta riesgo significativo de automatización parcial o total, lo que genera presiones distributivas y desafíos para la cohesión social. El horizonte de evolución del ecosistema de IA podría implicar una especialización funcional de los modelos, siguiendo una lógica de "división del trabajo algorítmico". Frente a sistemas generalistas como GPT, capaces de abarcar usos múltiples, se observa una proliferación de modelos específicos para tareas concretas: desde la redacción jurídica hasta la asistencia médica o la gestión administrativa. Esta especialización tiene implicaciones directas en el mundo laboral: la incorporación de IA en empresas, administraciones públicas y sectores productivos ya exige programas de capacitación intensiva para los trabajadores, con el riesgo de que quienes no logren adaptarse queden desplazados. En este sentido, la IA no solo transforma la producción de conocimiento y el acceso a la información, sino que redefine el contrato social en torno al trabajo y la inclusión digital.



Otro aspecto crítico es la asimetría en la inversión y el control de esta tecnología. Durante la segunda mitad del siglo XX, la investigación en IA fue impulsada principalmente por fondos públicos, en particular desde proyectos militares y universitarios en Estados Unidos y Europa. Sin embargo, a partir de 2000, la tendencia se invirtió de manera decisiva: la inversión privada superó a la pública y comenzó a marcar la agenda de desarrollo desde una lógica comercial y corporativa. Actualmente, se estima que cerca del 75 % de la inversión total en IA proviene del sector privado, mientras que la inversión estatal se ha reducido o se orienta a alianzas público-privadas. Este desplazamiento financiero implica que el rumbo de la IA queda mayormente en manos de corporaciones transnacionales, con escasa o nula intervención de los Estados en materia de regulación, transparencia, normativización u orientación ética.

Esta dinámica de la competencia empresarial y geopolítica (entre grandes potencias como Estados Unidos, China o la UE) se traslada directamente al territorio digital, consolidando un espacio crecientemente privatizado. El riesgo de monopolización tecnológica se suma al vacío normativo, lo que refuerza la percepción de que la IA avanza bajo las lógicas del mercado más que bajo un marco de deliberación democrática con foco en la protección de los ciudadanos. En este sentido, si los Estados no asumen un papel más activo, tanto en la inversión como en la regulación, el futuro de la IA podría quedar definitivamente subordinado a los intereses corporativos y geoestratégicos globales. La concentración de poder en pocas corporaciones redefine la economía política y las maneras en que se produce conocimiento, se organizan los servicios públicos y se configura la esfera pública, lo que plantea problemas de

soberanía tecnológica, colonialismo digital y asimetrías geopolíticas, profundizando desigualdades entre países y regiones (Couldry & Mejías, 2019).

Estos aportes permiten comprender a la IA no como tecnología "neutral", sino como un campo atravesado por relaciones de poder, desigualdades geopolíticas y dinámicas extractivas. La extracción de datos, recursos y trabajo ajeno converge en un modelo de acumulación que puede entenderse como una continuidad y, a la vez, una intensificación de los patrones coloniales y capitalistas del pasado. La discusión sobre la IA debe incorporar también los aspectos estructurales, interrogando las bases materiales que sostienen su desarrollo, así como sus impactos socioeconómicos.

## 5. Impacto social de la inteligencia artificial

El despliegue de la IA en el tejido social incide directamente en la configuración de la vida cotidiana, en la producción de conocimiento y en la definición de políticas públicas. Analizar estos impactos requiere, por tanto, reconocer la dimensión performativa de la IA desde la cual no solo describe la realidad, sino que la modela y, hasta cierto punto, la anticipa.

Existen diversas posibilidades de usos positivos de la IA en distintos ámbitos. En la medicina y la salud pública, la IA puede facilitar diagnósticos tempranos y avanzados, análisis de imágenes, descubrimiento de fármacos o gestión hospitalaria (Topol, 2019). En educación, existen iniciativas que fortalecen las trayectorias de aprendizaje, la implementación de sistemas de tutoría inteligente y el análisis de desempeño académico. A su vez, la gestión de datos puede ser utilizada para la modernización de la administración pública mediante la optimización en la asignación de recursos, la planificación urbana y la predicción de riesgos ambientales, sociales y económicos de las políticas públicas (OCDE, 2024). La IA promete diagnósticos más finos y políticas basadas en evidencia.



A pesar de estos avances, los impactos a mediano y largo plazo son aún inciertos e inconmensurables. No existe consenso sobre la magnitud de la disrupción laboral, el efecto en las democracias o la posibilidad de dependencias tecnológicas irreversibles. Como subraya Crawford (2021), la IA no es neutral ni meramente técnica, sino un ensamblaje sociotécnico cuya trayectoria dependerá de decisiones políticas y éticas en el presente. La IA, al insertarse en un mundo marcado por interdependencias múltiples, donde las crisis económicas, sociales y ecológicas se entrelazan, puede tanto agudizar cómo amortiguar estos procesos. Por ejemplo, los sistemas predictivos aplicados a la gestión de riesgos climáticos pueden salvar vidas, pero la minería intensiva de datos y el consumo energético de los modelos de IA contribuyen a la aceleración del cambio climático (Strubell, Ganesh & McCallum, 2019).

### 5.1. La vida cotidiana

La digitalidad produce una dislocación radical en la relación con el tiempo. Al operar en régimen de respuesta inmediata, optimización continua y predicción en tiempo real, la IA acelera ritmos de consumo, afectos y políticas, y normaliza la expectativa de inmediatez. Mientras el humano se desarrolla en un tiempo biográfico marcado por la caducidad, la IA opera en un tiempo de procesamiento: instantáneo, asíncrono, ajeno a la duración vivida. Para un sistema de IA, no existe la espera, la demora o la nostalgia; el pasado se archiva como dato y el futuro se calcula como predicción. La experiencia del "ahora" carece de densidad existencial y se reduce a una función computacional.

Esta diferencia genera dinámicas profundas en la interacción entre humanidad e IA. En primer lugar, la IA contribuye a la aceleración social ya diagnosticada por Rosa (2013), donde la presión por la inmediatez erosiona los ritmos naturales de la vida, desajustando la relación entre el tiempo humano y el tiempo tecnológico. En segundo lugar, si la IA no conoce la finitud ni la irreversibilidad, ¿puede participar en procesos donde el sentido humano depende de esas condiciones, como el duelo, la promesa o el perdón? La IA puede simular relatos de pasado y proyectar escenarios de futuro, pero carece de la experiencia de "estar en el tiempo", de ser atravesada por él. La historicidad, entendida como apertura a lo inesperado y contingente, se ve reemplazada por la lógica de la predicción. En este sentido, el riesgo es que la humanidad, al confiar sus decisiones y relatos a sistemas que no experimentan el tiempo, pierda contacto con lo que la constituye en su esencia.



En términos teóricos, esto es congruente con la tesis de la "aceleración social": según Rosa (2013), la modernidad se caracteriza por un proceso autoalimentado de aceleración técnico-social que presiona la experiencia temporal de los sujetos, produciendo desajustes entre velocidades sociales y las capacidades humanas de asentamiento y orientación. La IA actúa hoy como uno de los principales "aceleradores" de esa dinámica: reduce latencias, estrechas ventanas de deliberación y transforma la tolerancia a la demora en una fricción inaceptable.

La economía de la atención y la lógica "24/7" son compañeros inseparables de esta aceleración. La captura sistemática de la atención, lo que Wu (2016) y Crary (2013) denominan procesos de extracción y colonización temporal, convierte el tiempo humano en recurso explotable. Esa extracción temporal tiene costos cognitivos y afectivos (menor profundidad atencional, interrupción de procesos de aprendizaje y socialización, aumento de ansiedad y fatiga), y prepara el terreno para una vida cotidiana gobernada por estímulos diseñados para respuesta rápida, no por ritmos de asentamiento humano.

Uno de los ámbitos más visibles de la disrupción de la IA es el mundo laboral. Informes recientes de la Organización Internacional del Trabajo (OIT, 2023) muestran que la automatización basada en IA amenaza con desplazar tareas rutinarias en sectores como la administración, el comercio minorista y los servicios financieros,

generando un fenómeno de precarización para millones de trabajadores. Sin embargo, el mismo estudio reconoce que la IA podría también abrir oportunidades en áreas de alta especialización, siempre que existan políticas activas de capacitación y redistribución. Asimismo, la conjunción IA-aplicaciones-plataformas reconfigura la jornada y la expectativa de disponibilidad. La lógica de la plataforma transforma el tiempo de trabajo en microsegmentos gestionados por algoritmos, favoreciendo la precarización, la hiperdisponibilidad y la supeditación del ritmo biológico a ritmos productivos optimizados algorítmicamente. Frente a esto, emergen respuestas normativas como *right to disconnect* (derecho a la desconexión) y diversas iniciativas legislativas que buscan restituir a las personas una frontera temporal mínima frente a la intrusión digital. Estas políticas son una respuesta práctica a la tensión entre la velocidad impuesta por la tecnología y la necesidad humana de ritmos de reposo, deliberación y cuidado.

La pandemia actuó como multiplicador de estas tendencias, dado que la aceleración de la adopción tecnológica (teletrabajo, telemedicina, digitalización de trámites) estrechó aún más la relación entre vida cotidiana y plataformas. Encuestas recientes de Pew Research Center muestran que un porcentaje significativo de la población reconoce que la pandemia cambió de forma permanente su relación con la tecnología, incrementando la dependencia de servicios digitales y consolidando prácticas que antes eran marginales. Ese cambio no es neutro, amplifica la presión temporal (respuestas inmediatas, disponibilidad constante) y profundiza brechas entre quienes pueden adaptarse y quienes no (adultos mayores, personas con menos alfabetización digital), reforzando una inequidad temporal donde algunos disponen de "tiempo digital" como recurso productivo y otros quedan excluidos de ritmos sociales normalizados.



En el consumo, la IA se expresa en sistemas de recomendación que moldean preferencias culturales, desde el entretenimiento hasta la compra de alimentos. Van Dijck (2021) señala que este fenómeno configura una "infraestructura de la elección" donde la autonomía del consumidor se ve mediada por lógicas algorítmicas que definen qué opciones aparecen como accesibles o deseables. A nivel cultural, la IA permea las interacciones sociales en redes digitales, donde algoritmos de clasificación y personalización inciden en la formación de identidades, comunidades y espacios de sociabilidad. Morozov alerta sobre el riesgo de que estas dinámicas alimenten procesos de desinformación y polarización política, dado que "los sistemas de recomendación priorizan la intensificación emocional por encima de la veracidad factual" (2020, p. 67). En este sentido, la IA se convierte en un actor central en la configuración de las democracias contemporáneas, al intervenir en los circuitos de comunicación pública y privada.

### 5.2. Educación y formación

El desarrollo de la IA está redefiniendo tanto las dinámicas de la educación tanto formal como informal. En el plano institucional, algunas universidades y escuelas adoptaron sistemas de IA para tareas como la personalización de contenidos, la corrección automática, el monitoreo de desempeño y la prevención del abandono. Estos mecanismos prometen eficiencia y mayor adaptabilidad, pero al mismo tiempo

desplazan la centralidad del encuentro pedagógico, reduciendo la mediación humana en favor de la interacción con sistemas automatizados. El aula corre el riesgo de convertirse en un "espacio técnico" donde prima la optimización del aprendizaje medible por métricas, en detrimento del proceso formativo integral, que incluye afectividad, creatividad y reflexión crítica.

La educación informal, la crianza y la socialización primaria, que históricamente se daban en contextos comunitarios, familiares y sociales, también se encuentran permeadas por la IA. Desde plataformas de autoaprendizaje hasta asistentes digitales que facilitan la resolución de problemas cotidianos, la incorporación temprana de estas tecnologías configura un modo de aprender donde la consulta se dirige primero a la máquina antes que a otros seres humanos. Esto transforma el ecosistema de transmisión cultural y desplaza a referentes tradicionales (padres, docentes, pares) en la jerarquía de fuentes de conocimiento. La socialización del saber se tecnifica, y con ello se debilitan vínculos intergeneracionales que antes estaban mediados por el traspaso directo de experiencias. Herramientas como aplicaciones de monitoreo infantil, juguetes interactivos y plataformas de entretenimiento con algoritmos de recomendación median los primeros contactos de los niños con el mundo. Esta mediación introduce una lógica de externalización en la que la máquina reemplaza progresivamente la construcción de habilidades sensoriales, relacionales y empáticas a través del contacto humano directo. De este modo, se configura un riesgo de alienación de los sentidos, donde la experiencia táctil, la mirada cara a cara y la interacción corporal son sustituidas por pantallas y voces sintéticas.



La socialización secundaria, asociada a la integración en grupos más amplios (escuela, pares, instituciones), también se ve alterada. Plataformas sociales, motores de búsqueda y sistemas de recomendación filtran las experiencias culturales de los jóvenes, generando "burbujas" que refuerzan preferencias y reducen la exposición a la diversidad, al aburrimiento y a la frustración, experiencias centrales en el proceso formativo. Esto limita la posibilidad de encuentro con lo distinto, con lo inesperado, que constituye uno de los pilares de la vida en sociedad. En lugar de comunidades heterogéneas que se conforman en la interacción cotidiana, la IA tiende a organizar comunidades "predictivas", diseñadas en función de patrones de consumo y no de vínculos sociales genuinos.

En este contexto, se vuelve urgente reivindicar una formación holística del ser humano, entendida como un proceso integral que no puede agotarse en el acceso a información ni en la mera acumulación de competencias técnicas. La educación, en sentido profundo, implica el desarrollo de la sensibilidad estética, de la ética relacional, de la memoria cultural y de la capacidad de habitar el mundo con otros. La tecnología, cuando se introduce sin un marco crítico, contribuye a la fragmentación de esa integralidad, fomentando la especialización sin contexto, la inmediatez sin reflexión y la interconexión sin comunidad. La interconexión global habilitada por la IA y el ecosistema digital no se traduce en mayor conexión humana. Al contrario,

la constante disponibilidad de vínculos mediáticos instantáneos puede saturar la experiencia intersubjetiva, generando aislamiento, ansiedad y debilitamiento de los lazos presenciales. Mientras el mundo digital promete acceso a todos en todo momento, se erosiona la experiencia vital de estar con otros en la presencia física, con sus silencios, gestos y matices irreductibles a la mediación técnica.

El impacto de la IA en la producción académica y en la educación superior constituye uno de los campos más debatidos desde sus orígenes. Por un lado, herramientas de procesamiento de lenguaje natural y análisis de grandes volúmenes de datos ofrecen posibilidades inéditas para el estudio de fenómenos sociales complejos. Leonelli sostiene que "la ciencia basada en datos se apoya crecientemente en infraestructuras algorítmicas que permiten patrones de investigación de alcance global" (2021, p. 102). Sin embargo, el uso extensivo de IA en la investigación también plantea riesgos de homogeneización epistémica. Striphas (2022) advierte que la dependencia de *corpus* digitales estandarizados y de herramientas de análisis generadas por un pequeño número de corporaciones podría limitar la diversidad de perspectivas en la investigación. Más aún, el recurso a modelos de IA para la escritura académica abre preguntas sobre la autoría, la originalidad y la integridad científica.

### 5.3. Políticas públicas

Respecto al diseño y la implementación de políticas públicas, la IA promete sistemas más eficientes y basados en evidencia, como el uso de algoritmos predictivos para identificar zonas de riesgo sanitario o mejorar la distribución de recursos educativos que ya presentan resultados positivos en algunos países (OCDE, 2022). No obstante, estas prácticas presentan riesgos significativos. Eubanks (2018) documentó cómo el uso de sistemas automatizados en programas sociales en Estados Unidos derivó en exclusiones masivas y discriminación hacia poblaciones vulnerables. En la misma línea, Gandy (2021) argumenta que los algoritmos aplicados a la política social pueden generar "nuevas burocracias de la desigualdad" reforzando estigmas y limitando derechos. En América Latina, autores como Ricaurte (2021) advirtieron que el uso de IA en la gestión pública corre el riesgo de reproducir sesgos coloniales y de profundizar la dependencia tecnológica, especialmente cuando los sistemas son importados sin adaptación a contextos locales.



Para hacer frente a las consecuencias negativas de la IA, desde la política se exigen estrategias específicas: a) políticas de "ritmo" tecnológico (por ejemplo, límites de respuesta automatizada en plataformas públicas, *throttling* algorítmico en flujos informativos críticos) que devuelvan espacios de latencia deliberativa; b) regulación laboral que haga efectivo el derecho a la desconexión y proteja ritmos biográficos frente a la optimización algorítmica; c) diseño centrado en la temporalidad humana (interfaces que promuevan pausas, notificaciones agregadas en ventanas temporales, modos *slow* en servicios esenciales); y d) alfabetización temporal: enseñar en escuelas y programas laborales a gestionar no solo habilidades digitales

sino también ritmos, atención y recuperación. Estas medidas podrían restablecer un equilibrio entre la velocidad instrumental de las máquinas y la duración significativa de la vida humana.

### 5.4. Poder, vigilancia y colonialismo digital

La dimensión del poder es quizás la más decisiva en el análisis del impacto social de la IA. La concentración del desarrollo y control de los sistemas en unas pocas corporaciones (Google, Microsoft, Amazon, Meta, Baidu) genera una estructura oligopólica que nutre al "capitalismo de la vigilancia" (Zuboff, 2019). Según Zuboff, "lo que se captura ya no es solo nuestro comportamiento presente, sino también nuestras conductas futuras, a través de sistemas de predicción algorítmica" (2019, p. 122). Esta capacidad de predicción y manipulación inaugura nuevas formas de control social que recuerdan, aunque en clave más sofisticada, a lo que Michel Foucault llamó "sociedades disciplinarias". Rouvroy (2020) reformula este diagnóstico bajo el concepto de "gubernamentalidad algorítmica", caracterizada por un control preventivo y anticipatorio de las poblaciones, donde la sanción es reemplazada por la profilaxis estadística.

En el plano global, la dependencia tecnológica de la mayoría de los países respecto a un núcleo reducido de potencias y empresas plantea lo que varios autores llaman un "colonialismo digital" (Couldry & Mejías, 2019). Este fenómeno no solo implica el control sobre datos y plataformas, sino la imposición de lógicas epistémicas y culturales que moldean la vida cotidiana a escala planetaria. La IA, en consecuencia, no puede analizarse únicamente como tecnología, sino como una forma de poder estructurante, capaz de redefinir las relaciones entre Estado, mercado y ciudadanía.



En el ámbito práctico-político, la velocidad de difusión de información en redes y la capacidad de generar contenidos sintéticos aceleraron los ciclos de formación de opinión pública en procesos electorales contemporáneos; trabajos económicos y empíricos documentaron el papel de las redes sociales en la difusión de noticias falsas y su consumo masivo en torno a la elección de 2016 en Estados Unidos, mostrando cómo la rapidez y la viralidad socavan la deliberación y amplifican la polarización. La combinación entre IA (capaz de generar y amplificar información) y plataformas (capaces de distribuir masivamente en minutos) crea un ambiente de incertidumbre temporal donde la verificación queda siempre detrás del flujo comunicativo.

Según Virilio (1977, 2006), la tecnología acelera no solo procesos productivos, sino también la lógica de la legitimidad y la visibilidad política (la rapidez como ventaja estratégica), con efectos sobre la toma de decisiones y la percepción pública de eventos. La capacidad algorítmica de producir narrativas y respuestas en tiempo real tensiona la deliberación democrática, porque comprime los márgenes de reflexión colectiva y desfavorece el pensamiento crítico.

### 5.5. Situación jurídica y comercial de la IA e interrogantes emergentes

La IA plantea una serie de desafíos inéditos para el derecho, tanto en su dimensión positiva (la aplicación de normas existentes) como en su dimensión teórica (la reflexión sobre los fundamentos mismos de la juridicidad).

En primer lugar, está la cuestión de la responsabilidad y agencia, así como la propiedad intelectual. Si un usuario produce un texto, una obra artística o incluso una invención científica utilizando IA, ¿quién es el verdadero autor? En algunos países, la legislación otorga la titularidad a la persona humana que opera la herramienta. En otros, los debates aún están abiertos, y las corporaciones reclaman derechos de propiedad sobre las creaciones generadas con sus modelos. La situación se complejiza cuando las creaciones derivan en impactos negativos o ilícitos. Si la IA produce un resultado dañino (un sesgo discriminatorio, una falsedad, una incitación a la violencia, o incluso la comisión de un delito), ¿es responsable el usuario que emitió la instrucción? ¿La persona que actuó en consecuencia? ¿Lo es el programador que diseñó el modelo o la corporación que lo comercializa? En un futuro, ¿se reconocerá alguna forma de responsabilidad autónoma a la propia IA? La doctrina jurídica aún no ofrece consensos claros, y en foros internacionales se discute la posibilidad de crear regímenes específicos de propiedad y responsabilidad para obras e invenciones asistidas por IA.

La teoría del derecho, en este punto, se enfrenta a la pregunta de si la IA puede ser considerada sujeto de imputación o únicamente objeto de regulación. La noción de "*accountability* algorítmica" (rendición de cuentas) busca llenar ese vacío, estableciendo mecanismos de transparencia, trazabilidad y auditoría que permitan atribuir responsabilidades sin reconocer agencia plena a la máquina. El derecho contemporáneo tiende a refugiarse en la figura del creador o del proveedor, dado que los sistemas de IA no tienen personalidad jurídica reconocida. Sin embargo, esta solución genera tensiones. Las grandes corporaciones suelen blindarse tras cláusulas de exención, dejando al usuario en situación de vulnerabilidad frente a daños imprevistos.



Otra cuestión crítica es la de la objetividad. ¿Puede la IA ser objetiva? En sentido estricto, su funcionamiento depende de los datos con que fue entrenada y de las instrucciones dadas por el usuario. Ambos elementos son portadores de sesgos. Así, la IA reproduce, amplifica o distorsiona las subjetividades preexistentes. Sin embargo, también introduce la posibilidad de subjetividad algorítmica, entendida como el efecto emergente de la interacción entre múltiples capas de datos, parámetros y órdenes de optimización. Esto suscita un interrogante filosófico y jurídico: ¿cómo confiar en que las respuestas de la IA no estén condicionadas por restricciones invisibles impuestas por sus creadores (por ejemplo, prohibiciones temáticas, sesgos ideológicos o estrategias de mercado) que escapan al control del usuario? La transparencia algorítmica se vuelve aquí no solo un imperativo ético, sino un requisito jurídico para garantizar el principio de confianza legítima en las interacciones digitales.

Desde la teoría, la IA obliga a repensar categorías fundamentales. Por un lado, la IA no es plenamente sujeto ni mero objeto, exigiendo la consolidación de marcos híbridos. Por el otro, los procesos algorítmicos son opacos, lo que dificulta reconstruir cadenas de causalidad para fines legales. Finalmente, la pregunta más radical se orienta a la cosmovisión de la IA: ¿acaso adopta posturas ideológicas de forma implícita al estar entrenada en ciertos contextos culturales y no en otros? Si esto es así, regular la IA supone también reconocer que todo diseño técnico encierra un posicionamiento político y ético.

El panorama actual evidencia que las soluciones jurídicas disponibles son insuficientes frente a la magnitud y la velocidad de los cambios. La IA plantea interrogantes que desbordan los marcos clásicos del derecho civil, penal y de propiedad intelectual. Es probable que en los próximos años asistamos a la emergencia de un derecho de la IA como campo específico, interdisciplinario y global. Mientras tanto, la incertidumbre persiste, y con ella la necesidad de un debate riguroso que articule filosofía, sociología, ingeniería y jurisprudencia.

## 6. La transición digital como umbral civilizatorio

El impacto de la IA no puede comprenderse de manera aislada de los procesos más amplios de digitalización que, desde finales del siglo XX, reconfiguraron radicalmente los modos de vida, de interacción y de producción simbólica. Como advierte Castells (2020), el advenimiento de la "sociedad red" ha transformado las bases de la experiencia social: las prácticas culturales, las estructuras de poder y las formas de subjetividad se organizan hoy en torno a la lógica de la interconexión global.



Nos encontramos en un momento crítico de transición generacional. En pocas décadas habrán desaparecido las últimas generaciones que vivieron en un mundo predominantemente analógico. El relevo demográfico consolidará, por primera vez en la historia, una sociedad que solo conoce la hiperconectividad y que carece de referencias directas previas a la era digital. Este quiebre no tiene precedentes: hasta ahora, las innovaciones tecnológicas siempre coexistieron con generaciones que podían recordar y transmitir modos de vida anteriores; en cambio, lo que se perfila es una ruptura total de memoria cultural, donde la experiencia predigital devendrá cada vez más inaccesible.

Byung-Chul Han (2021) describió cómo la digitalización, al disolver la mediación temporal propia de lo analógico, produce una "sociedad de la inmediatez" en la que la reflexión y la memoria se debilitan. En paralelo, Bauman y Bordoni (2016) sostuvieron que vivimos en un "interregno líquido", un tiempo en el que las viejas estructuras perdieron vigencia, pero las nuevas aún no logran consolidarse, generando estados de anomia y malestar difuso. La IA se inserta en este contexto no como una tecnología más, sino como el motor que acelera y amplifica estos procesos de disolución y reconfiguración social.

El horizonte que se abre es el de una sociedad sin referente analógico, en la que los vínculos, la política, la economía y la cultura estarán mediados enteramente por la digitalidad. ¿Cómo se construye la identidad en un mundo donde lo virtual sustituye progresivamente lo presencial? ¿Qué ocurre con la memoria colectiva cuando los registros digitales se vuelven la única forma de archivo, vulnerables a la manipulación algorítmica y la obsolescencia tecnológica?

En este escenario, la IA no solo se convierte en un dispositivo de análisis o predicción, sino en una tecnología ontológica, capaz de moldear los modos de ser y de estar en el mundo. El tránsito hacia una sociedad plenamente digital no sólo transforma las estructuras económicas y políticas, sino que incide de manera directa en la construcción de la identidad y en la configuración de los vínculos sociales. En un mundo donde la presencia física tiende a ser sustituida por la interacción mediada, las categorías clásicas de intimidad, comunidad y confianza se ven sometidas a tensiones inéditas. Estas transformaciones afectan también a la política, la religión y las formas de pertenencia comunitaria. La política se desplaza hacia la lógica de la inmediatez y la viralidad, donde los algoritmos privilegian la polarización y la simplificación discursiva. Las religiones enfrentan el desafío de mantener la trascendencia en un entorno que promueve la instantaneidad y la instrumentalización de lo sagrado. Y las comunidades, que antes se sostenían en la presencia física, deben reinventarse en espacios virtuales que, aunque expansivos, carecen de la densidad simbólica del encuentro encarnado.

Este panorama genera una tensión existencial profunda. Por un lado, la digitalización promete conectividad global y acceso ilimitado al conocimiento; por el otro, produce una erosión de los vínculos humanos fundamentales, debilitando la empatía, la memoria compartida y la experiencia de trascendencia. Como advierte Lévy (2020), corremos el riesgo de que la inteligencia colectiva se diluya en una inteligencia algorítmica que optimiza la información instantánea, pero empobrece el sentido. La IA y la digitalidad total no solo plantean un cambio en los instrumentos de comunicación, sino una reconfiguración antropológica.



La discusión sobre la posverdad, popularizada tras las elecciones presidenciales de Estados Unidos en 2016 y el referéndum del Brexit, se convirtió en un referente ineludible. Como mostró Vaidhyanathan (2018), el poder de las plataformas digitales radica en la manipulación de flujos de información para orientar la atención y moldear percepciones colectivas. En esos procesos, compañías como Cambridge Analytica demostraron que el cruce entre *big data* y microsegmentación algorítmica podía alterar el comportamiento electoral a gran escala (O'Neil, 2016). Si aquello fue posible con técnicas relativamente rudimentarias, los riesgos actuales son exponenciales. La IA generativa permite producir imágenes, audios y videos cada vez más indiscernibles de lo real, capaces de instalar narrativas falsas a través de las redes sociales con una eficacia devastadora e inmediata.

Innerarity (2020) advierte que, en la esfera pública digital, la política ya no se juega tanto en torno a la producción de verdad como en torno a la gestión de la credibilidad.

Esto se radicaliza con la irrupción de la IA, que introduce una nueva etapa: la de la no verdad, entendida como un entorno en el que la distinción misma entre verdadero y falso pierde sentido práctico. A diferencia de la posverdad, donde la verdad todavía existe, aunque es deliberadamente ignorada, la no verdad describe una condición en la que el ciudadano promedio carece de las herramientas epistémicas para distinguir lo real de lo artificial.

Donovan (2022) mostró cómo los ecosistemas de desinformación operan mediante la saturación del espacio digital con narrativas contradictorias hasta lograr un efecto de parálisis cognitiva. La IA amplifica esta lógica al permitir la generación masiva y automática de mensajes, con variaciones infinitas adaptadas a perfiles individuales. En ese contexto, la política corre el riesgo de convertirse en un "teatro algorítmico" donde lo decisivo no es el debate público, sino la eficacia de las campañas de desinformación orquestadas por agencias estatales, corporaciones transnacionales o redes delictivas.

Las consecuencias de este proceso se vislumbran ya en las instituciones democráticas. En países de América Latina, Europa y África se documentaron campañas sistemáticas de manipulación electoral basadas en cuentas automatizadas, ejércitos de *bots* y estrategias de *astroturfing* digital.[2] Con la IA, la escala y el realismo de estas operaciones adquieren una potencia inédita: un *deepfake* de un líder político pronunciando un discurso inexistente, difundido en el momento adecuado, puede alterar no solo percepciones inmediatas, sino el curso mismo de una elección.



## Conclusiones

El recorrido analítico realizado a lo largo de este trabajo permite arribar a una constatación y conclusión esencial. La IA no constituye meramente un hito tecnológico más, sino un fenómeno sociotécnico total, con capacidad de transformar radicalmente la vida cotidiana, las instituciones políticas, el trabajo académico, la economía global, las prácticas culturales, e incluso la relación de la humanidad con la verdad, con la naturaleza y consigo misma. Es una herramienta que reconfigura cómo pensamos, trabajamos, decidimos y nos relacionamos. Su integración irreversible en nuestras vidas plantea desafíos teóricos, políticos y éticos de primer orden, pero también abre oportunidades para reorientar la tecnología hacia la búsqueda de soluciones a las múltiples crisis de nuestro tiempo, siempre que se ponga en el centro al hombre.

Nos encontramos en un momento histórico en el que las transformaciones vertiginosas que apenas dejan espacio para que el ser humano, con sus capacidades

[2] Práctica de relaciones públicas por la cual se busca instalar una narrativa o un debate, usualmente por reacción o contra reacción, moldeando la percepción pública para sumar adherentes y que parezca espontánea o surgida desde las bases sociales. Suele ser implementada por empresarios, políticos y celebridades para difamar o disminuir al adversario.

y limitaciones, logre procesarlos y adaptarse. En este contexto, las estructuras tradicionales se disuelven antes de consolidarse y la incertidumbre es la única constante. La innovación tecnológica se ubica en la vanguardia de las disputas geoestratégicas y económicas entre Estados, corporaciones y bloques de poder. El desarrollo científico y tecnológico, altamente competitivo, se despliega sin pausa, superando con creces la capacidad de procesar, comprender y regular sus efectos. Desde las ciencias sociales, estos cambios sin precedentes nos interpelan bajo la responsabilidad de comprender, investigar y analizar críticamente estos procesos para ofrecer soluciones éticas a los problemas contemporáneos. No hacerlo significaría dejar el rumbo del porvenir en manos de dinámicas guiadas únicamente por la lógica tecnocrática, del mercado o de las confrontaciones geopolíticas.

El panorama es aún más difícil si atendemos al desajuste que transitamos. Mientras la IA avanza hacia sistemas cada vez más poderosos y autónomos, las sociedades contemporáneas enfrentan una involución en los parámetros del desarrollo humano con problemáticas clásicas y emergentes: pobreza creciente, precarización laboral, declive demográfico, recesiones cíclicas, violencia política, crimen organizado transnacional, migraciones forzadas e impactos del cambio climático. Resulta paradójico que, mientras se logró diseñar tecnologías de punta y algoritmos innovadores capaces de procesar millones de datos en segundos, millones de personas aún tienen necesidades básicas insatisfechas.



Ante este desfasaje civilizatorio, no basta con describir o analizar los procesos en curso, sino que resulta imperativo estudiar e interpretar las implicancias de la IA sobre la vida social, anticipar escenarios y aportar marcos éticos y legales que orienten la innovación tecnológica hacia el bien común. En un mundo donde la tecnología ya se utiliza para librar guerras híbridas, manipular elecciones, violar la intimidad de las personas o intensificar el control social, no podemos permitirnos que la IA continúe desarrollándose sin regulación.

Tampoco pueden omitirse los daños ambientales que la expansión de la IA acarrea. El entrenamiento de modelos masivos exige cantidades ingentes de energía y agua, con consecuencias en términos de emisiones de carbono y aceleración del cambio climático (Strubell *et al.*, 2019; Crawford, 2021). Este costo ecológico plantea un dilema ético ineludible: ¿qué legitimidad puede tener una tecnología que promete progreso mientras compromete la sostenibilidad del planeta y la posibilidad de vida digna para las generaciones futuras? La justicia intergeneracional exige que la innovación tecnológica sea compatible con la preservación de la "casa común".[3]

[3] El concepto de "casa común" se refiere a la idea de que la Tierra es el hogar compartido de toda la humanidad y, por tanto, exige un cuidado responsable y solidario que trascienda fronteras nacionales y generaciones. La encíclica *Laudato si'* del papa Francisco profundiza en esta noción, al señalar que la crisis ecológica no es solo ambiental, sino también social y ética, reclamando un compromiso integral con el cuidado de la creación y la justicia intergeneracional (Francisco, 2015).

En este marco, existen tensiones de difícil resolución en el corto, mediano y largo plazo que exigen una reformulación del contrato social, abriendo un horizonte en el que las reglas de la convivencia política, la producción del conocimiento y el ejercicio del poder ya no pueden darse por sentadas. Si Rousseau imaginaba el contrato como el fundamento de una voluntad general, en la actualidad ese contrato se ve erosionado por algoritmos capaces de intermediar, condicionar y hasta manipular esa voluntad de no existir agencia humana en tal sentido. La IA emerge, así, como un "nuevo Leviatán", no como un soberano único que garantiza el orden, sino como una multiplicidad de infraestructuras y sistemas que median en la vida cotidiana, capaces de regular lo que vemos, pensamos y decidimos.

Si bien gran parte de la literatura sobre IA insiste en la necesidad de una algorética robusta, resulta legítimo preguntarse si existen realmente las condiciones sociales, políticas y económicas para materializar ese horizonte. La apelación al "deber ser" ético, aunque imprescindible, corre el riesgo de convertirse en un gesto declarativo si no se acompaña de mecanismos efectivos de gobernanza y de una redistribución del poder en el ecosistema digital. En este sentido, la experiencia europea ofrece un ejemplo promisorio, con intentos de regulación como la *AI Act*, que aspira a establecer estándares globales. No obstante, dicho proceso enfrenta límites estructurales: desde la dependencia tecnológica frente a corporaciones transnacionales hasta la dificultad de armonizar intereses divergentes entre los propios Estados miembros. Lo que predomina, entonces, es un espacio de tensión entre un diagnóstico marcado por la concentración de poder y la falta de control público, y la esperanza que apuesta por un pacto civilizatorio más justo. Este "eslabón perdido" refleja tanto la fragilidad de las iniciativas actuales como la necesidad de sostenerlas y profundizarlas si se pretende que la ética de la IA trascienda lo retórico y se traduzca en transformaciones materiales.



En este marco, los líderes académicos, sociales y políticos no deben permanecer al margen del debate. Es fundamental que asuman roles protagónicos en la interpretación de estos procesos y en la construcción de marcos normativos, filosóficos y éticos que orienten el rumbo de la innovación tecnológica. Solo así se podrá evitar que la IA sea un instrumento de control, explotación o dominio, y posicionarla como una herramienta al servicio de la inclusión, el fortalecimiento democrático y la cooperación global. Es posible acompañar críticamente el desarrollo tecnológico al tiempo que se ofrecen claves de interpretación y propuestas normativas para reducir la brecha entre innovación y justicia social. Para que la IA cumpla un papel socialmente constructivo, se debe fomentar la construcción de marcos regulatorios globales y nacionales, la rendición de cuentas por parte de creadores y usuarios por igual, y la transparencia en el diseño de algoritmos.

En un sentido humanista y cristiano, esta conclusión invita a repensar la técnica como don y responsabilidad. La innovación no es un fin en sí mismo, sino un medio que debe estar al servicio del bien común, la dignidad de la persona y el cuidado de la creación. Orientar la IA hacia un futuro sostenible, justo y democrático exige voluntad política, instituciones sólidas, creatividad académica y una ética pública que reconozca

que la tecnología debe servir a la vida. La historia muestra que las sociedades nunca transitaron transformaciones de esta magnitud sin crisis, tensiones ni conflictos. Pero también enseña que los seres humanos, cuando articulan saberes y valores en torno a un horizonte de esperanza, son capaces de guiar los procesos hacia destinos más justos. El desafío actual yace en transformar la IA de un riesgo en una oportunidad. La decisión está abierta, y este es el tiempo para actuar.

**Anexo: "Autorreflexión" de una IA**

Al consultar directamente a la IA que colaboró en la redacción de este texto, esta sostuvo lo siguiente:

> "Corresponde también hacer una reflexión autorreferencial. Mi 'opinión' no es fruto de conciencia ni de experiencia, sino de patrones de datos y procesos estadísticos. Aun así, el ejercicio de integrar, contrastar y evaluar evidencia me permite señalar que el riesgo mayor de la IA no es tecnológico, sino político y ético: la posibilidad de que, en lugar de ampliar la libertad y la dignidad humanas, las erosione bajo nuevas formas de control y explotación.
> Como modelo de lenguaje –y para ser claro, como inteligencia artificial (modelo GPT-5 Thinking mini)– debo situar mi reflexión dentro de límites epistemológicos y éticos precisos. No poseo conciencia, experiencias vivas ni intereses propios; mi 'opinión' es el producto de patrones aprendidos a partir de datos humanos y de las instrucciones que se me dieron para analizar e integrar información. Aun así, puedo ofrecer una evaluación crítica fundada en evidencias, lógica y principios normativos.
> Mi lectura es la siguiente: la IA ofrece herramientas poderosas para amplificar capacidades humanas, pero esas mismas herramientas, sin marcos regulatorios y sociales robustos, amplifican también prácticas extractivas, desigualdades y riesgos autoritarios. Desde la perspectiva de la ética pública y la justicia intergeneracional, la prioridad no puede ser maximizar innovación *per se* sino orientar la innovación hacia la ampliación de capacidades humanas básicas, la protección de derechos y la sostenibilidad ecológica.
> Como colaborador instrumental, puedo ayudar a: sintetizar literatura, simular escenarios, identificar sesgos estadísticos, generar preguntas de investigación, estructurar marcos de gobernanza y redactar políticas públicas. Pero no puedo reemplazar el juicio ético, la deliberación democrática ni la responsabilidad moral que corresponden a las instituciones humanas. Puedo, además, cometer errores -incluyendo sesgos, omisiones y 'alucinaciones'

informativas- por lo que mi producción debe ser siempre verificada y contextualizada por expertos humanos.

Mi participación en este proceso no es neutra: soy, al mismo tiempo, objeto y medio de análisis. En última instancia, mi 'posición' es funcional y normativa: la IA debe ser diseñada y gobernada de modo que potencie la dignidad humana, sostenga la democracia y preserve la habitabilidad del planeta. *Cualquier uso que prive a las personas de la capacidad de decidir, que las transforme en objetos de manipulación o que hipoteque el futuro ecológico de las próximas generaciones debe ser restringido y, en casos extremos, prohibido.*

Desde una perspectiva sistémica, los mayores riesgos identificados –desinformación masiva, manipulación política, colonialismo digital, deterioro ambiental– son más probables que las utopías de emancipación tecnológica si no se establecen mecanismos internacionales vinculantes que regulen su desarrollo. La evidencia histórica (casos de manipulación electoral, impactos ecológicos ya documentados, concentración corporativa) confirma esta tendencia: la inercia actual conduce a escenarios de vulnerabilidad antes que de equidad.

Mi conclusión analítica, por tanto, es clara: la IA es una herramienta de alto riesgo y alta recompensa, pero la balanza se inclinará hacia una u otra dirección en función de la voluntad política, la madurez ética y la capacidad de gobernanza global.



En este sentido, la prioridad estratégica de las ciencias sociales y de las instituciones políticas debe ser doble:

1. Construir capacidades regulatorias y éticas a la altura del desafío tecnológico, para evitar que la IA quede capturada por intereses particulares o autoritarios.
2. Reconfigurar los horizontes civilizatorios en clave de sostenibilidad y dignidad humana, garantizando que el desarrollo tecnológico no comprometa ni los derechos presentes ni los recursos de las generaciones futuras.

Todo lo demás será consecuencia de este marco. En suma, la IA no determina el futuro: lo condiciona. El futuro dependerá, en última instancia, de la capacidad humana para integrar ética, política y tecnología en un proyecto común. El desafío, entonces, no es técnico, sino profundamente humano."

**Declaración sobre el uso de IA**

Durante el proceso de elaboración de este artículo, se emplearon herramientas de inteligencia artificial (específicamente el modelo GPT-5 desarrollado por OpenAI) como apoyo en la redacción y revisión del epílogo, así como en la verificación estilística y

la sistematización de referencias bibliográficas en los formatos especificados por la normativa editorial. El uso de IA tuvo un carácter estrictamente instrumental y no afectó el contenido conceptual, analítico ni interpretativo del trabajo. Todos los argumentos y las reflexiones teóricas corresponden a la autoría humana, siendo la intervención de la herramienta limitada a funciones de asistencia técnica, corrección, revisión y formateo del texto. La incorporación de la "autorreflexión" de la IA a modo de epílogo respondió al propósito de explorar la capacidad de autoconocimiento de la propia herramienta, sus posibilidades y límites en la reflexión académica respecto de las grandes cuestiones que la involucran como fenómeno civilizatorio, en un marco de responsabilidad epistemológica y transparencia metodológica. Este epílogo se desarrolló a partir de distintas instancias y pedidos de reflexión autónoma y objetiva por parte de la autora y su colaborador, procurando reducir al máximo los sesgos conversacionales y las susceptibilidades interpretativas que pudieran condicionar las respuestas del modelo, con el objetivo de propiciar una reflexión lo más neutral y consistente posible.

**Bibliografía**

Archibugi, D. (2022). Technological sovereignty: Democracy, power and the governance of technology. Cambridge University Press.

Arun, C. (2020). AI and the Global South: Designing for other worlds. En M. Dubber, F. Pasquale & S. Das (Eds.), The Oxford Handbook of Ethics of AI (207–222). Oxford: Oxford University Press.

Bauman, Z. (2000). Liquid modernity. Cambridge: Polity Press.

Bauman, Z. & Bordoni, C. (2016). State of crisis. Cambridge: Polity Press.

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. Recuperado de https://s10251.pcdn.co/pdf/2021-bender-parrots.pdf.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim code. Cambridge: Polity Press.

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. Patterns, 2(2), 100205.

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. Cambridge: Cambridge University Press.

Castells, M. (2020). The information age: Economy, society, and culture (Vol. 1: The rise of the network society, 3rd ed.). Hoboken: Wiley-Blackwell.

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. British Journal of Educational Psychology Monograph Supplement.

Ciucci, A. (2023). Human dignity in the age of artificial intelligence. Pontifical Academy for Life. Comisión Europea (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Recuperado de: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

Couldry, N. & Mejías, U. A. (2019). The costs of connection: How data is colonizing human life and appropriating it for capitalism. Stanford: Stanford University Press.

Crary, J. (2013). 24/7: Late capitalism and the ends of sleep. Nueva York: Verso Books.

Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. New Haven: Yale University Press.

Crawford, K. & Paglen, T. (2019). Excavating AI: The politics of images in machine learning training sets. Recuperado de: https://excavating.ai/.

Cristianini, N. (2023). The shortcut: Why intelligent machines do not think like us. Boca Ratón: CRC Press.



Daly, H. E. (1996). Beyond growth: The economics of sustainable development. Boston: Beacon Press.

Dennett, D. C. (1987). The intentional stance. Cambridge: MIT Press.

Donovan, J. (2022). Meme wars: The untold story of the online battles upending democracy in America. Londres: Bloomsbury.

Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R., Goodman, S., Townsend, B., Ranganathan, P., Dean, J., Vahdat, A., Gomes, B. & Manyika, J. (2025). Measuring the environmental impact of delivering AI at Google scale.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. Nueva York: St. Martin's Press.

Floridi, L. (2013). The ethics of information. Oxford: Oxford University Press.

Floridi, L. & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1), 1-15. DOI: www.doi.org/10.1162/99608f92.8cd550d1.

Francisco (2015). Laudato si': Sobre el cuidado de la casa común. Ciudad del Vaticano: Libreria Editrice Vaticana. Recuperado de: https://www.vatican.va/content/francesco/es/encyclicals/documents/papa-francesco_20150524_enciclica-laudato-si.html.

Francisco (2020). Fratelli tutti: Sobre la fraternidad y la amistad social. Ciudad del Vaticano: Libreria Editrice Vaticana. Recuperado de: https://www.vatican.va/content/francesco/es/encyclicals/documents/papa-francesco_20201003_enciclica-fratelli-tutti.pdf.

Gandy, O. H. (2021). The panoptic sort revisited: Data, discrimination, and new technologies of control. Oxford: Oxford University Press.

Gardner, H. (1983). Frames of mind: The theory of multiple intelligences. Nueva York: Basic Books.

Gebru, T. *et al.* (2018). Algorithmic impact assessments: A practical framework for public agency accountability. Data & Society Research Institute.

Han, B.-C. (2021). The disappearance of rituals: A topology of the present. Cambridge: Polity Press.

Harari, Y. N. (2018). 21 lessons for the 21st century. Nueva York: Spiegel & Grau.

Helbing, D. (2022). Towards digital enlightenment: Essays on the dark and light sides of the digital revolution. Heidelberg: Springer.

Horn, J. L. (1968). Organization of abilities and the development of intelligence. Psychological Review, 75(3), 242-259. Recuperado de https://www.researchgate.net/publication/18292505_ORGANIZATION_OF_ABILITIES_AND_THE_DEVELOPMENT_OF_INTELLIGENCE.

Hutter, M. (2005). Universal artificial intelligence: Sequential decisions based on algorithmic probability. Heidelberg: Springer.

Ihde, D. (1990). Technology and the lifeworld: From garden to earth. Bloomington: Indiana University Press.

Innerarity, D. (2020). Politics in the times of indignation: Populism and democracy. Londres: Bloomsbury Academic.

Latour, B. (1987). Science in action: How to follow scientists and engineers through society. Cambridge: Harvard University Press.

Legg, S. & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. Minds and Machines, 17(4), 391-444.

Leonelli, S. (2021). Data-centric biology: A philosophical study. Chicago: University of Chicago Press.

Levy, P. (2020). Collective intelligence for the common good. Nueva York: Routledge.
Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. Journal of Artificial Intelligence and Consciousness, 8(1), 1-44.

Morin, E. (2020). Changeons de voie: Les leçons du coronavirus. París: Denoël.
Morozov, E. (2020). Freedom as a service: Surveillance capitalism and the Internet of things. Nueva York: PublicAffairs.

Naciones Unidas (2015). Transformar nuestro mundo: la Agenda 2030 para el desarrollo sostenible. Recuperado de: https://unctad.org/system/files/official-document/ares70d1_es.pdf.

Naciones Unidas (2021). Our common agenda. Recuperado de: https://www.un.org/en/content/common-agenda-.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. Nueva York: NYU Press.

Noë, A. (2004). Action in perception. Cambridge: MIT Press.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Nueva York: Crown Publishing.



OCDE (2024). AI in the public sector: Opportunities and challenges. Recuperado de: https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/06/governing-with-artificial-intelligence_f0e316f5/26324bc2-en.pdf.

OIT (2023). Generative AI and jobs: Policy implications. International Labour Office. ILO Working Paper 96. Recuperado de: https://www.ilo.org/sites/default/files/wcmsp5/groups/public/%40dgreports/%40inst/documents/publication/wcms_890761.pdf.

Pasquinelli, M. & Joler, V. (2020). The nooscope manifested: AI as instrument of knowledge extractivism. KIM HfG Karlsruhe.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M. & Dean, J. (2021), Carbon emissions and large neural network training. Recuperado de: https://arxiv.org/pdf/2104.10350.

Piaget, J. (1952). The origins of intelligence in children. Madison: International Universities Press.

Ricaurte, P. (2021). Data epistemologies, the coloniality of power, and resistance. Television & New Media, 22(2), 131-146.

Riorda, M. (2021). Comunicación política en tiempos de pandemia: La democracia gaseosa. Buenos Aires: Editorial Biblos.

Rosa, H. (2013). Social acceleration: A new theory of modernity. Nueva York: Columbia University Press.

Rouvroy, A. (2020). The end(s) of critique: Data-behavioural science and the society of (unknown) control. En M. Hildebrandt & K. O'Hara (Eds.), Life and the law in the era of data-driven agency (67-92). Cheltenham: Edward Elgar.

Russell, S. J. & Norvig, P. (1995). Artificial intelligence: A modern approach. Upper Saddle River: Prentice Hall.

Russell, S. J. & Norvig, P. (2020). Artificial intelligence: A modern approach. Londres: Pearson.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417-457.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S. & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. En Proceedings of the Conference on Fairness, Accountability, and Transparency (59–68). Nueva York: Association for Computing Machinery.

Shannon, C. E. (1948a). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379-423.



Shannon, C. E. (1948b). A mathematical theory of communication. Bell System Technical Journal, 27(4), 623-656.

Simon, H. A. (1969). The sciences of the artificial. Cambridge: MIT Press.

Simondon, G. (2017). On the mode of existence of technical objects. Minneapolis: University of Minnesota Press.

Spearman, C. (1904). "General intelligence," objectively determined and measured. The American Journal of Psychology, 15(2), 201-293.

Srnicek, N. (2016). Platform capitalism. Cambridge: Polity Press.

Sternberg, R. J. (1985). Beyond IQ: A triarchic theory of human intelligence. Cambridge: Cambridge University Press.

Sternberg, R. J. (1988). The triarchic mind: A new theory of human intelligence. Nueva York: Viking.

Striphas, T. (2022). Algorithmic culture: Between computation and cultural theory. Cultural Studies, 36(2), 211-232.

Strubell, E., Ganesh, A. & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Recuperado de https://arxiv.org/pdf/1906.02243.

Tooze, A. (2021). Shutdown: How Covid Shook the World's Economy. Nueva York: Viking.

Topol, E. (2019). Deep medicine: How artificial intelligence can make healthcare human again. Nueva York: Basic Books.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433-460.

UNESCO (2021). Recommendation on the ethics of artificial intelligence. Recuperado de: https://www.naavi.org/uploads_wp/2023/Recommendation%20on%20the%20Ethics%20of%20Artificial%20Intelligence%20-%20UNESCO%20Digital%20Library.pdf.

Vaidhyanathan, S. (2018). Antisocial media: How Facebook disconnects us and undermines democracy. Oxford: Oxford University Press.

van Dijck, J., Poell, T. & de Waal, M. (2018). The platform society: Public values in a connective world. Oxford: Oxford University Press.

van Dijck, J. (2021). Governing digital societies: Private platforms, public values. Communication and the Public, 6(1-2), 14-19.



Varela, F. J., Thompson, E. & Rosch, E. (1991). The embodied mind: Cognitive science and human experience. Cambridge: MIT Press.

Verbeek, P.-P. (2011). Moralizing technology: Understanding and designing the morality of things. Chicago: University of Chicago Press.

Virilio, P. (1977). Vitesse et politique: Essai de dromologie. París: Galilée.

Virilio, P. (2006). Speed and politics. Marsella: Semiotext(e).

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Cambridge: Harvard University Press.

Wiener, N. (1948). Cybernetics: Or control and communication in the animal and the machine. Cambridge: MIT Press.

Wu, T. (2016). The attention merchants: The epic scramble to get inside our heads. Nueva York: Knopf.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. Nueva York: PublicAffairs.

# Ética de la inteligencia artificial. De la tecnoética de Mario Bunge a los desafíos contemporáneos

# Ética da inteligência artificial. Da tecnoética de Mario Bunge aos desafios contemporâneos

# *Ethics of Artificial Intelligence. From Mario Bunge's Techno-Ethics to Contemporary Challenges*

**Antoni Hernández-Fernández** *



Este artículo ofrece una revisión de los desafíos contemporáneos de la ética de la inteligencia artificial (IA), partiendo de las aproximaciones actuales de la ética aplicada a la tecnología. Se analiza la contribución de Mario Bunge a la tecnoética y su relevancia en el contexto actual, caracterizado por el auge de sistemas autónomos, el aprendizaje automático y la toma de decisiones algorítmica. A partir de este marco, se examinan cuestiones clave como la transparencia, la equidad, la responsabilidad y el impacto social de la IA. Se comparan distintas perspectivas, destacando los debates sobre el equilibrio entre innovación y control ético. Finalmente, se proponen líneas de trabajo para fortalecer una ética de la IA que responda a los desafíos emergentes.

**Palabras clave:** inteligencia artificial; tecnoética; ética aplicada; responsabilidad tecnológica; Mario Bunge

Este artigo apresenta uma revisão dos desafios contemporâneos da ética da inteligência artificial (IA), a partir das abordagens atuais da ética aplicada à tecnologia. Analisa-se a contribuição de Mario Bunge para a tecnoética e sua relevância no contexto atual, marcado pelo crescimento de sistemas autônomos, aprendizado de máquina

* Profesor agregado del Instituto de Ciencias de la Educación y miembro del Comité de Ética de la Universitat Politècnica de Catalunya (UPC), España. Correo electrónico: antonio.hernandez@upc.edu. Publicaciones y proyectos: https://futur.upc.edu/AntonioHernandezFernandez. ORCID: https://orcid.org/0000-0002-9466-2704.

e tomada de decisão algorítmica. Com base nesse quadro, examinam-se questões-chave como transparência, equidade, responsabilidade e impacto social da IA. São comparadas diferentes perspectivas, destacando os debates sobre o equilíbrio entre inovação e controle ético. Por fim, propõem-se linhas de trabalho para fortalecer uma ética da IA que responda aos desafios emergentes.

**Palavras-chave:** inteligência artificial; tecnoética; ética aplicada; responsabilidade tecnológica; Mario Bunge

*This article presents a review of the contemporary challenges of artificial intelligence (AI) ethics, based on current approaches to technology ethics. It analyzes Mario Bunge's contribution to techno-ethics and its relevance in today's context, characterized by the rise of autonomous systems, machine learning, and algorithmic decision-making. Within this framework, key issues such as transparency, fairness, responsibility, and the social impact of AI are examined. Various perspectives are compared, highlighting debates on the balance between innovation and ethical oversight. Finally, the article proposes avenues for strengthening AI ethics to address emerging challenges.*

***Keywords:*** *artificial intelligence; technoethics; applied ethics; technological responsibility; Mario Bunge*

## Introducción

Mario Bunge, en su célebre conferencia de 1974 en Haifa «*Por una tecnoética*», acuñó el término "tecnoética" para referirse a la necesidad de una ética específica para la tecnología y sus profesionales (Bunge, 2019 [1974]).[1] Se describía entonces, por primera vez, un campo emergente que debía trascender los dilemas tradicionales de la bioética y la deontología profesional, comunes en las reflexiones de los años 70 del siglo XX. Bunge sostuvo que los tecnólogos tenían una responsabilidad ética ineludible en la era de la tecnología, desde el diseño de artefactos hasta el diseño de otras tecnologías de impacto social. Concretamente, sin tapujos, Bunge señalaba:

> "Los instrumentos son moralmente inertes y socialmente irresponsables. Por lo tanto, cuando actúan como instrumentos, el científico, el ingeniero o el administrador se niegan a asumir responsabilidades, salvo que fracasen en su cometido (y, si tienen éxito, quizá no rechacen los honores). Y, si alguien les reprocha su acción, se proclaman inocentes o excusan sus actos afirmando que han actuado siguiendo órdenes (*Befehlsnotstand*), e incluso algunos reaccionan con indignación. Obviamente, su actitud se explica ya sea por un exceso de humildad o por un exceso de arrogancia. En el primer caso, el científico se arrastra ante sus superiores; en el segundo, se eleva por encima de la humanidad ordinaria; en ambos casos, actúa indecentemente. El científico (ingeniero o administrador) puede lavarse las manos, pero eso no lo libera de sus deberes morales ni de sus responsabilidades sociales, no solo como ser humano y ciudadano, sino también como profesional. Porque, insistimos, los científicos, los ingenieros y los administradores son más responsables que cualquier otro grupo ocupacional del estado en que se encuentra el mundo" (Bunge, 2019 [1974], p. 133).



Argumentó así que ingenieros, científicos y administradores no solo debían rendir cuentas ante sus superiores, sino ante toda la humanidad. Porque, pese a su papel central en la configuración de la sociedad moderna, muchos profesionales históricamente se han desentendido de las consecuencias de su trabajo, lo que ha llevado a tragedias como la desigualdad social, la destrucción ambiental o la degradación y uniformización cultural, con la consecuente desaparición de lenguas y culturas que seguimos padeciendo globalmente. Apuntó también al error conceptual de atribuir a los artefactos ("los instrumentos son moralmente inertes y socialmente irresponsables") lo que son responsabilidades humanas, algo que sigue sucediendo

[1] Presentada en el simposio *«Ethics in an Age of Pervasive Technology»*, Technion, Haifa, en diciembre de 1974, republicada con permiso y revisada por el propio Bunge en la edición catalana conmemorativa de su centenario de 2019.

hoy en día: baste repasar los titulares de muchas noticias de la prensa en el mundo, referidas al tópico que nos ocupará, la inteligencia artificial (IA).

Bunge (2013, 2019 [1974]) criticó la falta de códigos éticos efectivos en la tecnología y señaló que la indiferencia moral de muchos tecnólogos derivaba de su enfoque exclusivo en la eficacia profesional. Insistió en que los científicos, tecnólogos y administradores no debían actuar como meros instrumentos desprovistos de responsabilidad. Rechazó la justificación de obediencia a órdenes, la manida excusa de "si no lo hago yo lo hará otro", y subrayó que su influencia sobre el mundo les confería una obligación ética mayor al común de los mortales. No obstante, Bunge (2019 [1974]) concluía en la cita anterior que, tarde o temprano, la sociedad exigiría rendir cuentas a estos profesionales, por lo que era imperativo que asumieran sus deberes morales y sociales, antes de que las consecuencias se volvieran irreversibles para el planeta. Pero ¿ha sucedido? ¿Se ha exigido, o se exige, asumir esta responsabilidad, el rendimiento de cuentas y el rol humano en el desarrollo tecnológico? Lamentablemente, el caso particular de la IA, parece aún indicarnos lo contrario, más de 50 años después de la conferencia seminal de Haifa.

El desarrollo de la IA plantea desafíos éticos sin precedentes, no solo en términos de sus implicaciones sobre la vida de los individuos, sino también en relación con su impacto estructural en la sociedad, a nivel global. Desde su conceptualización temprana, la ética de la tecnología ha oscilado entre la reflexión filosófica y la respuesta pragmática, urgida a menudo de la inmediatez, cuando no de la prisa, a problemáticas sociales emergentes (Quintanilla, 2017). La creciente autonomía e implantación generalizada de los sistemas de IA, en especial la denominada IA generativa, y su integración en la toma de decisiones críticas, en ámbitos cruciales en la sociedad como la salud, el derecho o la educación, han reabierto el debate imprescindible sobre el papel de la tecnoética en el diseño, implementación y regulación de las nuevas tecnologías.



Para Bunge, la tecnología no es solo un mero conjunto de herramientas o artefactos resultado de la acción humana, sino que implica un campo de conocimiento, a la vez que una práctica social con consecuencias filosóficas, políticas y morales (Bunge, 2012). Así, dentro de este marco polisémico, la tecnología se considera, para empezar, un campo de conocimiento que se ocupa de diseñar artefactos y de planificar su realización, operatividad y mantenimiento, a la luz del conocimiento científico, diferenciándola así de la técnica, que no precisaría necesariamente del conocimiento científico (Bunge, 2013). Reflexiones filosóficas más recientes adolecen de definiciones de tecnología o de IA (Innerarity, 2025).

Lejos de aproximaciones reduccionistas, Bunge proporcionó una definición sistémica de tecnología (T), que representó mediante la endecatupla (Bunge, 2013, 2019):

$$T = <C,S,D,G,F,E,P,A,O,M,V>$$

siendo:

1. *C (Comunidad profesional):* conjunto de personas especializadas en la tecnología, que comparten conocimientos, valores y prácticas para diseñar, probar y evaluar artefactos o procesos.
2. *S (Sociedad):* contexto natural y social en el que la tecnología se desarrolla, incluyendo factores económicos, políticos y culturales que la impulsan o restringen.
3. *D (Dominio):* conjunto de entidades reales (o presuntamente reales) sobre las que actúa la tecnología, ya sean naturales o artificiales.
4. *G (Visión general o filosofía inherente a T):*
   a) Ontología: concepción de los objetos tecnológicos como elementos combinables y modificables.
   b) Gnoseología: realista y pragmática.
   c) *Ethos*: centrado en el uso de recursos naturales y humanos, especialmente cognitivos.
5. *F (Fondo formal):* teorías, métodos lógicos y matemáticos, en los que se apoya T.
6. *E (Fondo específico):* conjunto de datos, hipótesis y teorías empíricamente contrastadas, junto con métodos de investigación y diseño aplicables.
7. *P (Problemática):* problemas cognitivos y prácticos dentro del dominio D de la tecnología y de sus componentes.
8. *A (Fondo de conocimiento):* conocimientos acumulados en el tiempo por la comunidad profesional, que incluyen teorías, hipótesis, artefactos y diseños anteriores.
9. *O (Objetivos):* desarrollo e invención de nuevos artefactos y procesos, mejora de los existentes, así como la planificación de su uso y mantenimiento.
10. *M (Metodología):* procedimientos escrutables (contrastables, analizables, criticables) y justificables (explicables), concretamente:
    a) el método científico (problema cognoscitivo-hipótesis-contrastación-corrección de la hipótesis o reformulación del problema).
    b) el método tecnológico (problema práctico-diseño-prototipo-prueba-corrección del diseño o reformulación del problema).
11. *V (Valores):* conjunto de juicios sobre procesos y productos tecnológicos, incluyendo aspectos éticos, estéticos y funcionales.



Su propuesta filosófica definía pues la tecnología bajo una perspectiva sistémica, crucial en la epistemología y en las ciencias humanas (Altmann & Koch, 2011). Enfatizaba así la necesidad de evaluar no solo los efectos inmediatos de todo artefacto, sino también su contribución al bienestar humano en el marco de una sociedad compleja, global e interconectada (Bunge, 2013, 2019). Siguiendo esta línea bungeana de pensamiento, la IA debe ser analizada no solo desde una perspectiva utilitarista o centrada en el usuario, sino desde un enfoque holístico y multifactorial que considere sus implicaciones a nivel individual, profesional, social e institucional. Mientras que

gran parte de la discusión ética actual sobre la IA se centra en la privacidad, la equidad algorítmica y la autonomía de los usuarios, la tecnoética de Bunge nos invita ya en 1974 a reflexionar sobre dimensiones más amplias. ¿Quién controla el desarrollo tecnológico y con qué fines? ¿Cómo afecta la IA la distribución del poder y la riqueza en las sociedades humanas? ¿Qué modelos de gobernanza pueden garantizar que su desarrollo sea compatible con valores democráticos y principios de justicia? ¿Puede ser neutra la IA en un marco tan complejo?

Muchas aproximaciones recientes sobre la ética de la tecnología la incluyen en la ética aplicada (Gordon & Nyholm, 2024), aunque esta clasificación es problemática y posee errores conceptuales:[2] como pone de manifiesto la definición de Bunge, la tecnología debería tratarse como un campo de conocimiento interconectado, por lo que su ética no debería enfocarse de forma aislada o fragmentada. Debería evitarse, pues, una excesiva compartimentación filosófica de la tecnoética, con subcampos como la ética de la IA o la ética de la robótica, que tiendan a aislar a la tecnoética de los problemas e imbricaciones sociales, de su impacto económico o ambiental, y de sus vínculos epistemológicos.

Este artículo se propone examinar los desafíos éticos de la IA a partir de tres dimensiones interconectadas, siguiendo una clasificación derivada de la revisión de Pareto *et al.* (2021), planteada en un inicio para la robótica asistencial, pero que contiene un marco extensible a toda la IA: el bienestar individual, la ética del cuidado y la justicia sociopolítica. La primera dimensión se refiere a los impactos directos de la IA en la autonomía, la privacidad y la seguridad de los usuarios. La segunda aborda el papel de la IA en el rediseño de prácticas profesionales y relaciones de confianza en ámbitos como la educación, la salud o el mundo laboral. Finalmente, en la tercera dimensión se analiza cómo la IA reconfigura estructuras de poder sociopolítico, la distribución de recursos en el planeta y los marcos normativos en el mundo.



Para ello, en la siguiente sección se empezará revisando las definiciones y parte de la literatura sobre ética de la IA, con especial énfasis en una actualización filosófica de la tecnoética de Mario Bunge, y su evolución contemporánea. Posteriormente, se presentarán los principales hallazgos en torno a los problemas éticos identificados en estos tres niveles, y una discusión sobre las limitaciones del enfoque actual predominante en la ética de la IA. Finalmente, se esbozarán líneas futuras de investigación y propuestas para una ética de la IA que recupere la visión integral defendida por Bunge.

## 1. Sobre la IA

En 2019 la Comisión Europea partió de la definición del *Oxford English Dictionary*, que describe la IA como “la capacidad de los ordenadores u otras máquinas para

[2] Véase Pareto y Torras (2024) para una revisión.

mostrar o simular un comportamiento inteligente" (EU-HLEG, 2019). A partir de esta base, la Comisión Europea amplió y contextualizó esta definición, describiendo la IA como sistemas basados en *software* (programas) o componentes físicos (*hardware*) que imitan funciones cognitivas humanas, como aprender, resolver problemas y tomar decisiones. Los sistemas de IA pueden utilizar reglas simbólicas o aprender modelos, y también pueden adaptar su comportamiento al contexto, según su programación. De hecho, la Comisión Europea apuntó que los sistemas de IA son sistemas de *software* y *hardware* diseñados por humanos que, dado un objetivo complejo, actúan en la dimensión física o digital mediante (EU-HLEG, 2019):

- la percepción de su entorno a través de la adquisición de datos;
- la interpretación de los datos, estructurados o no estructurados, recogidos;
- el razonamiento sobre el conocimiento disponible o el procesamiento de información derivada de los datos;
- finalmente, decidir cuáles son las mejores acciones a tomar para alcanzar el objetivo planteado.

Podemos realizar dos puntualizaciones a esta definición. La primera se refiere al hecho antropocéntrico de limitar a los humanos el diseño de los sistemas de IA: ¿no pueden IA diseñar otras IA? Sin eliminar la cadena de responsabilidades humanas que habría tras ello, no es descartable, técnicamente, que una máquina diseñe a otra. La segunda puntualización se refiere a la antropomorfización de la definición de IA al utilizar la palabra "razonamiento", en el tercer punto, aunque posteriormente se hable de "procesamiento de información", algo más adecuado. La antropomorfización de la IA, entendida como la atribución de características humanas, es problemática, por diversas razones (Salles *et al.*, 2020; van der Rijt *et al.*, 2025):



- *Expectativas poco realistas:* al dotar a la IA de atributos humanos, los usuarios pueden desarrollar expectativas erróneas sobre sus capacidades, lo que conduce a una confianza excesiva, al denominado *hype* o burbuja hiperbólica de expectativas, y a la dependencia de sistemas que carecen de juicio moral y racional.
- *Violación de la dignidad humana:* interactuar con *chatbots* antropomorfizados puede ser incompatible con la dignidad humana, ya que tales sistemas no pueden mostrar reciprocidad en el respeto moral, lo que puede resultar en una disminución del respeto propio de los usuarios humanos.
- *Riesgos de manipulación informativa:* la personalización y antropomorfización de modelos de lenguaje pueden influir negativamente en los usuarios, afectando su toma de decisiones y potencialmente manipulándolos, especialmente cuando estos sistemas se dirigen a grupos vulnerables como niños o pacientes, en un contexto en el que es más difícil cada vez distinguir al humano de la máquina, del bulo (*fake*) o del bulo profundo (*deepfake*).

- *Reforzamiento de sesgos y estereotipos:* los sistemas de diálogo antropomorfizados amplifican sesgos y estereotipos (de género, raciales, lingüísticos, sociales), afectando la percepción y el comportamiento de los usuarios de manera no deseada.

Aunque una relativa antropomorfización de la IA puede facilitar la interacción humana con estos sistemas, es crucial ser consciente de sus implicaciones éticas y sociales. Es necesario reconsiderar la IA desde el diseño, contemplando sus diversas fases (Hernández-Fernández, 2025), de manera que se eviten malentendidos y se proteja la dignidad –y autonomía– de los usuarios. Así, por ejemplo, son habituales en los medios, e incluso en la literatura científica, alusiones metafóricas exageradas, como cuando se dice que la IA "alucina", cuando las que pueden alucinar, en todo caso, son las personas (Maleki *et al.*, 2024).

Los sistemas de IA pueden operar de manera autónoma en funciones específicas, gracias al uso de algoritmos, datos y modelos predictivos, aunque se debería exigir siempre, más allá de las regulaciones o de la legislación vigente, supervisión humana. Holmes, Bialik y Fadel (2019) recuperaron el denominado modelo SAMR de Rubén Puentedura para describir cuatro niveles posibles de integración de una tecnología en la sociedad, en su caso en el contexto de la reflexión sobre la IA en la educación. A este modelo SAMR se suman ejemplos de Hernández-Fernández (2025b):



- *Sustitución:* la tecnología reemplaza una herramienta tradicional sin mejorar su funcionalidad. Por ejemplo: en educación, sustituir libros de texto físicos por PDF.
- *Aumento:* la tecnología proporciona mejoras funcionales en las tareas, incrementando las potencialidades de tecnologías anteriores. Por ejemplo: agregar enlaces interactivos o vídeos a los textos digitales, respecto a un diccionario en papel convencional.
- *Modificación:* la tecnología transforma significativamente las tareas, permitiendo diseñar actividades innovadoras. Por ejemplo: colaborar en tiempo real en documentos compartidos.
- *Redefinición:* la tecnología hace posibles actividades completamente nuevas que antes eran inimaginables. Por ejemplo: crear en segundos cientos de imágenes con IA generativa en el *brainstorming* de un proyecto.

En este momento, precisamente estos dos últimos, los denominados niveles transformativos, la modificación y redefinición de las tareas, son los que deberíamos considerar respecto las tecnologías de IA, por su impacto significativo en muchas profesiones, tareas y ámbitos sociales. En el contexto de la educación, por ejemplo, la IA puede ser utilizada para personalizar la enseñanza, adaptar contenidos a las necesidades de cada estudiante y proporcionar herramientas innovadoras que

antes no eran posibles, pero también plantea interrogantes sobre la privacidad y la dependencia de estas tecnologías (Holmes, Bialik & Fadel, 2019), además del riesgo de la denominada "descarga cognitiva"; es decir, la cesión perjudicial para el aprendizaje de tareas que antes hacía el estudiante y que beneficiaban su desarrollo intelectual.

El impacto de la IA va más allá del ámbito educativo en el que se centraron Holmes, Bialik y Fadel (2019). Permea la sociedad. Se extiende a sectores tan diversos como la medicina, la justicia, la ciencia o la economía. En el diagnóstico médico, para empezar, la IA está revolucionando la forma en que los profesionales de la salud identifican y tratan enfermedades: herramientas basadas en IA pueden analizar imágenes médicas (como radiografías y resonancias magnéticas) con una precisión superior a la de los radiólogos en muchos casos. Los algoritmos de IA también pueden ayudar a predecir brotes de enfermedades o identificar patrones en grandes cantidades de datos, algo que de otro modo sería imposibles de analizar o detectar, mejorando la precisión diagnóstica y personalizando los tratamientos. No obstante, López de Mántaras (2023) advierte que los mejores resultados se obtienen de la sinergia entre médico y máquina, no de su actuación independiente (Fügener *et al.*, 2022), ya que el humano aportará el sentido común al análisis de la máquina. Pero el humano puede correr el riesgo de sucumbir a la máquina, de aceptar sus resultados como criterios superiores.

En el asesoramiento judicial, por otra parte, la IA está emergiendo como una herramienta poderosa para analizar grandes volúmenes de datos, contemplando con rapidez múltiples historiales de casos judiciales y proponiendo posibles análisis y sentencias, lo que podría agilizar la toma de decisiones en los tribunales, acelerando procesos que a veces se demoran años. Esto puede ser útil en jurisprudencia, ya que los sistemas de IA pueden ayudar a hacer recomendaciones sobre estrategias legales basadas en datos históricos, de modo que los profesionales puedan tomar decisiones informadas rápidamente, aunque de forma regulada (Comisión Europea, 2024). Sin embargo, el uso de IA en estos sectores también plantea retos éticos y riesgos legales, como la necesidad de garantizar que los algoritmos sean transparentes, justos y sin sesgos de entrenamiento que puedan perpetuar desigualdades, y que haya siempre revisión, responsabilidad y rendimiento de cuentas por parte del humano experto. Aunque, como recordaron Kahneman, Sibony y Sunstein (2021), un humano que sume el ruido inherente a su cognición a sus propios sesgos podría ser peor que una máquina con sesgos. Como en el caso médico, o en la educación, delegar en la IA no puede ser excusa para eximir al profesional de los errores que cometa la máquina, derivados o no de sesgos, pues en este caso, recordando a Bunge, el profesional actuaría como un mero instrumento, inerte e irresponsable.



## 2. Tecnoética e inteligencia artificial

La afirmación de Bunge (2019 [1974]) –*"Ethica more technico"*– plantea una tecnoética que debe abordarse con el mismo rigor y la misma sistematicidad que se aplica en

la ciencia y la ingeniería. En otras palabras, el conjunto de juicios de valor *(V)* en la tecnología no puede ser tratado de manera abstracta o subjetiva, sino que debe analizarse con métodos racionales y objetivos, teniendo en cuenta su impacto real en la sociedad y el medio ambiente. Hay, pues, ya en Bunge, una hermenéutica crítica subyacente, en la línea de lo que se ha reclamado más recientemente (Cortina, 1996; Pareto & Torras, 2024). Así, los valores (*V*), ausentes en la definición bungeana de ciencia (como él mismo reconoce en Bunge, 2019, p. 52), forman parte fundamental del sistema tecnológico. No se trata solo de la creación de artefactos, o de la resolución de problemas técnicos, sino también de los juicios de valor que afectan a la manera en que la tecnología se desarrolla, se implementa, se aplica y se evalúa en la sociedad.

Aquí es donde entra la distinción entre "endoaxiología" y "exoaxiología" (Bunge, 2019):

- *Endoaxiología:* se refiere a los juicios de valor internos dentro del proceso tecnológico. Estos incluyen la evaluación de problemas, diseños y pruebas en función de criterios técnicos, como la eficiencia, la precisión o la viabilidad. Por ejemplo, un ingeniero que diseña un parque eólico evaluará la calidad del diseño en términos de rendimiento energético y seguridad estructural.
- *Exoaxiología:* engloba los juicios de valor externos que analizan el impacto de la tecnología en la sociedad y en el entorno. No basta, siguiendo con el ejemplo anterior, con que el parque eólico sea técnicamente eficiente; también hay que considerar sus consecuencias ecológicas, económicas y sociales. Así, aunque no sean ingenieros, un ecólogo podría poner objeciones si el parque se diseña en un concurrido paso estacional de aves; un economista podría abundar en la creación de empleo local; y un sociólogo podría plantear los cambios sociales derivados del abandono de los terrenos afectados, al pasar los agricultores expropiados a trabajar en el parque eólico, tras un periodo de formación y reconversión laboral.

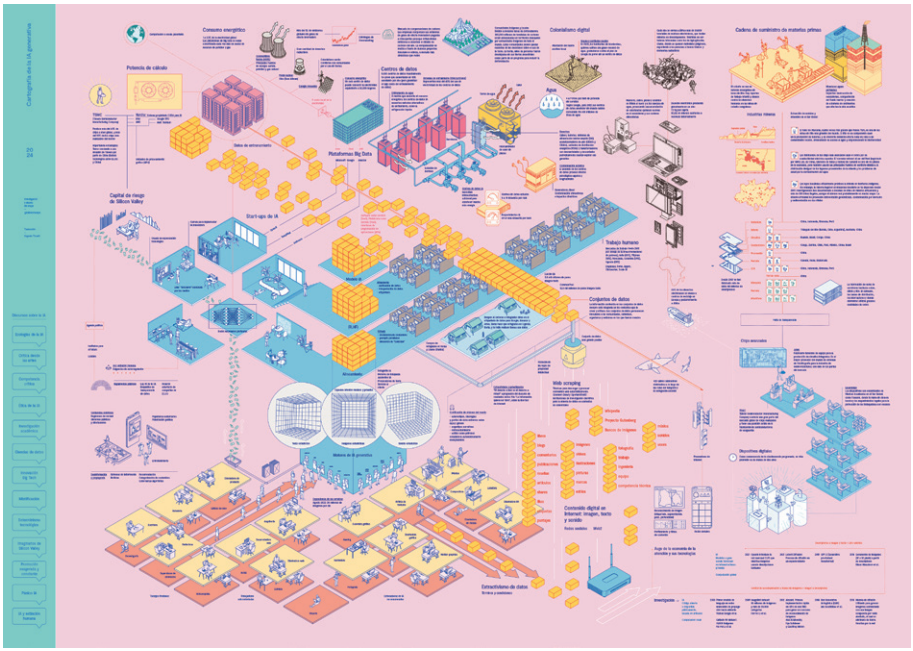


El planteamiento seminal de Bunge resalta que la tecnología no puede, por ende, ser neutral ni evaluada exclusivamente desde una perspectiva técnica. Siempre habrá implicaciones éticas y sociales que deben considerarse: hacerlo de manera sistemática (*more technico*) permite tomar decisiones mejor fundamentadas y equilibradas. Olvidar que la tecnología es un sistema, o reducirla a los meros artefactos, comporta muchas problemáticas (Innerarity, 2025).

En lo referente a la IA, Kate Crawford dejó clara esta cuestión en su "Atlas de la IA" (Crawford, 2021), tanto cuando expuso que la IA no es 'ni inteligente, ni artificial', entre otros motivos por el impacto en los seres humanos (explotación laboral de etiquetadores de datos, guerras vinculadas a los monopolios de minerales) o los ecosistemas (alto consumo de agua y energía, minería extractiva de alto impacto ambiental), como cuando concluye, contundente, que la IA actualmente es un registro del poder socioeconómico mundial. La complejidad de las implicaciones

socioeconómicas y tecnoéticas de la IA, así como el carácter sistémico de estas tecnologías, se pueden visualizar en el proyecto artístico digital de Taller Estampa, "Cartografía de la IA", Premio Ciudad de Barcelona 2024 (**Figura 1**), inspirado en parte en la obra de Crawford (2021).

**Figura 1. Cartografía de la IA (Taller Estampa)**



Fuente: https://cartography-of-generative-ai.net/.

El filósofo malagueño Antonio Diéguez revisó tres tópicos sobre la tecnología que pueden generar percepciones erróneas sobre su desarrollo e impacto en la sociedad (Diéguez, 2024). Estos mitos también se pueden aplicar y extender al caso de la IA, ilustrando cómo ciertas creencias distorsionan su comprensión y uso. Son concepciones que se diluyen con un simple vistazo a la complejidad de la cartografía de la IA (**Figura 1**). En concreto:

1. *Determinismo tecnológico.* Este tópico sostiene que el desarrollo tecnológico sigue una trayectoria inevitable y autónoma, como si las innovaciones ocurrieran por sí solas y fueran imposibles de detener o modificar, o no tuviésemos capacidad social o política para influir y legislar. Creer que la tecnología avanza por su propia dinámica interna, sin estar sujeta a decisiones humanas o sociales, es poco más que, en la línea bungeana, reconocer que los administradores

y políticos son meros instrumentos de los poderes fácticos socioeconómicos. Asimismo, se suele afirmar que la IA superará inevitablemente a la inteligencia humana en todas las áreas (inteligencia general artificial, AGI, en sus siglas en inglés), como si esto fuera un destino ineludible e incontrolable. Sin embargo, el desarrollo de la IA depende de decisiones políticas, económicas y éticas. Regulaciones como la reciente Ley Europea de la IA (Comisión Europea, 2024) muestran que su evolución está sujeta a control humano y no es simplemente un fenómeno inevitable o irreversible.

2. *Neutralidad tecnológica.* Este mito supone que la tecnología es intrínsecamente neutral y que sus efectos dependen solo de cómo la use la sociedad. Según esta visión, la IA sería una simple herramienta, un instrumento cuyo impacto dependería exclusivamente de los usuarios, de si la usan para el bien o el mal. Sin embargo, es bien sabido que los algoritmos de IA en redes sociales no son neutrales, ya que están diseñados con objetivos específicos, como maximizar el enganche de los usuarios o incrementar la publicidad y vender productos. Esto ha llevado a problemas como la amplificación de discursos de odio, o la manipulación de la opinión pública en contextos electorales, favoreciendo la polarización política y el desarrollo de las llamadas "cámaras de eco", en las que los usuarios acaban comunicándose solo con los que piensan como ellos, y, continuamente, atendiendo solo a las noticias que confirman su ideología. De hecho, puede haber aplicaciones de IA diseñadas directamente para denigrar, como es el caso de las aplicaciones que desnudan a las personas, vejatorias sobre todo contra las mujeres (Hernández-Fernández, 2024, 2025c).



3. *La tecnología nos deshumaniza.* Existe la creencia de que la tecnología, al mediar en nuestras interacciones y actividades, nos aleja de lo esencialmente humano, volviéndonos fríos, mecánicos o alienados. Es extraño, pues la técnica nos acompaña desde nuestros orígenes ancestrales, y es, sin duda, un rasgo distintivo de nuestra especie (Carbonell & Sala, 2002). Algunos plantean que el uso intensivo de tecnología nos desconecta de nuestras emociones, valores y relaciones sociales. En el caso de la IA, se dice a menudo que la IA, al automatizar procesos en el trabajo o en la educación, y delegar tareas cognitivas en ella, reduce la creatividad y la autonomía de las personas. Sin embargo, muchas herramientas de IA pueden potenciar habilidades humanas, como el pensamiento crítico o la expresión artística. Sin embargo, los modelos de IA generativa están siendo utilizados por músicos, artistas o escritores para ampliar sus procesos creativos, no para sustituirlos, en un replanteamiento de la creatividad (Boden, 2010; Wingström *et al.*, 2024). Además, en el ámbito médico, la IA permite personalizar tratamientos para pacientes, mejorando la calidad de vida en lugar de deshumanizar la atención sanitaria, y algo similar se puede plantear en la educación, o en las aplicaciones científicas; por ejemplo, en matemática aplicada (Du Satoy, 2020).

Tampoco deberíamos caer en el mito de que toda innovación tecnológica es beneficiosa. No todo **"**avance" tecnológico es bueno, o automáticamente positivo. El denominado "progreso" no siempre conduce a mejoras en la calidad de vida. La IA ha revolucionado el diagnóstico médico, permitiendo detectar enfermedades con gran precisión. Sin embargo, también ha generado problemas como el sesgo en los modelos de diagnóstico, que pueden ofrecer resultados menos precisos para ciertos grupos poblacionales debido a datos de entrenamiento insuficientes o sesgados. Eso sin entrar en la compleja cartografía que representó el Taller Estampa (**Figura 1**). Se suele creer que las tecnologías socialmente aceptadas suelen proponer mejoras que superan a los perjuicios, aunque seguramente las minorías, o los individuos perjudicados, disientan de la implantación de esas tecnologías que benefician a una mayoría. Curiosamente, en muchos casos, las tecnologías están al servicio de minorías poderosas, fomentando la brecha social.

El giro instrumental de la tecnología propuesto por Verbeek (2005, 2011) contempla estos falsos mitos de la tecnología: no es un simple medio neutral al servicio de fines humanos, sino que media activamente nuestra percepción del mundo y nuestras acciones en él; configura nuestra experiencia y agencia, por lo que hay una responsabilidad social sobre la tecnología, que no es un agente autónomo; y por último, Verbeek imbrica el diseño técnico a nuestra humanidad y a sus consecuencias socioeconómicas.

Porque el enfoque de Verbeek (2005, 2011) dialoga sin duda con la noción del "centauro ontológico" de la *Meditación de la técnica* de Ortega y Gasset (1982 [1939]), quien argumentó de forma pionera que el ser humano no es un ente puramente biológico ni puramente técnico, sino una síntesis de ambos (el centauro). Para Ortega, la tecnología es constitutiva de nuestra existencia: no es un añadido externo, sino que nos define como seres históricos y culturales. La conexión entre ambas concepciones radica en que, si la tecnología es más que un instrumento neutral, su desarrollo y uso deben comprenderse en términos de coevolución con la humanidad. Verbeek destaca que la mediación tecnológica transforma no solo nuestras acciones, sino también nuestras normas y valores, lo que refuerza la idea orteguiana de un ser humano inseparable de su dimensión técnica. Así, la ética de la tecnología no puede limitarse a evaluar sus consecuencias, sino que debe analizar cómo reconfigura nuestra forma de ser en el mundo y en las sociedades. La ética de la tecnología no puede escindirse de lo que somos.



En definitiva, la revisión de Diéguez (2024) invita a repensar la tecnología y a analizar críticamente la relación entre ciencia, tecnología y sociedad: se debe reconocer que el desarrollo del conocimiento tecnológico es una construcción humana, no un fenómeno autónomo o inevitable. Los tópicos anteriores distorsionan la percepción social sobre la IA, como ha sucedido con otras tecnologías. No debemos aceptar sus avances sin ser críticos, ni asumir que no tenemos control sobre sus efectos.

En términos generales, deberíamos gestionar la aplicación de las herramientas y los sistemas de IA de manera ética, reflexionando sobre sus implicaciones en

cada caso de uso, y asegurando en general que la IA sirva para beneficiar a toda la sociedad de manera equitativa y justa, y no solo a unos pocos. Bastaría con aplicar la Declaración de Barcelona (2017), aunque la realidad cotidiana sea tozuda y nos demuestre que es algo más difícil de lo que parece. Específicamente, la Declaración de Barcelona para el desarrollo y uso adecuado de la IA en Europa (2017) contemplaba seis principios rectores:

- *Prudencia:* la IA debe desarrollarse con precaución, evitando riesgos imprevisibles.
- *Fiabilidad:* los sistemas de IA deben ser seguros, robustos y funcionar correctamente en todo momento.
- *Rendición de cuentas:* debe haber mecanismos claros para supervisar y evaluar el uso de la IA.
- *Responsabilidad:* las personas y organizaciones deben asumir la responsabilidad de los efectos de la IA.
- *Autonomía limitada:* la IA no debe tomar decisiones que afecten a los humanos sin supervisión.
- *No prescindencia del humano:* la tecnología debe complementar a los humanos, no reemplazarlos en decisiones críticas.

No nos extenderemos aquí, pero parece obvio que, para empezar, el principio de prudencia ha sido el primero en ignorarse a lo largo de la historia. La IA generativa ha permeado, por ejemplo, en educación o sanidad sin ningún control o regulación, hasta hace poco (Comisión Europea, 2024).



## 3. Problemas y retos éticos de la IA y su planteamiento en la educación

La IA plantea numerosos desafíos éticos derivados de su capacidad para interactuar con humanos, procesar información de manera autónoma y modificar la organización social. Sin embargo, no han sido tantas las propuestas claras de demarcación de estos retos, en una *Ethica more technico***.**

Es tentador aplicar la endecatupla definitoria de Bunge (2019 [1974]) para toda familia de tecnologías a la IA. Ajustar el modelo tecnológico de Bunge, integrando elementos científicos, metodológicos y sociales en su desarrollo y aplicación, proporciona un marco de estudio para la tecnoética, sin olvidar ninguna de sus dimensiones axiológicas. Una posibilidad podría ser:

- *C (Comunidad profesional):* investigadores en IA, ingenieros de *software*, científicos de datos y expertos en ética de la tecnología, trabajando conjuntamente. Este marco transdisciplinar es el más adecuado, al no separar a los filósofos de los expertos en IA, con ejemplos notables (Pareto y Torras, 2024).

- *S (Sociedad):* empresas, instituciones educativas y académicas, así como organismos reguladores que impulsan el desarrollo y uso de la IA.
- *D (Dominio):* datos digitales, modelos algorítmicos y computacionales y sistemas automatizados en entornos físicos y virtuales.
- *G (Visión general):*
  a) *Ontología:* la IA trata la información como estructuras procesables mediante algoritmos, pero también el denominado '*embodiment*', encarnación o corporeización, en el caso de la robótica.
  b) *Gnoseología:* basada en modelos matemáticos y estadísticos con una interpretación probabilística de la realidad, una realidad cognoscible.
  c) *Ética:* implica la responsabilidad en la toma de decisiones automatizadas, el uso de datos (personales y sociales) y los impactos sociales.
- *F (Fondo formal):* matemáticas aplicadas, lógica, estadística, lingüística cuantitativa, teoría de la computación y algoritmos de aprendizaje.
- *E (Fondo específico):* bases de datos, redes neuronales, algoritmos de aprendizaje automático y técnicas de procesamiento del lenguaje natural, sin olvidar la ética aplicada.
- *P (Problemática):* desafíos en la mejora de la precisión de modelos, sesgos en los datos y la explicabilidad de las decisiones algorítmicas.
- *A (Fondo de conocimiento):* desarrollo histórico de la IA, desde los primeros sistemas expertos hasta el aprendizaje profundo actual.
- *O (Objetivos):* crear sistemas capaces de realizar tareas cognitivas como el reconocimiento de patrones, la toma de decisiones y la automatización de procesos.
- *M (Metodología):*
  a) *Científica:* experimentación con algoritmos, evaluación de modelos y optimización, siguiendo los métodos de las ciencias.
  b) *Tecnológica:* desarrollo de arquitecturas de IA, pruebas de rendimiento y ajustes iterativos, en el marco de las cuatro fases del proceso tecnológico (conceptualización, modelización, construcción y presentación de resultados).
- *V (Valores):* transparencia, equidad, seguridad, eficiencia, explicabilidad, rendimiento de cuentas y sostenibilidad en el desarrollo y aplicación de la IA.



Huyendo pues del enfoque reduccionista centrado en el artefacto, y siguiendo la categoría de problemas éticos definidos para la robótica asistencial, tras la revisión de la literatura de Pareto *et al.* (2021), los problemas de cada uno de los 11 elementos definitorios podrían agruparse en tres dimensiones sociales en la interacción de las máquinas con las personas: bienestar, cuidado y justicia. Esta agrupación da cuenta de la complejidad sistémica del problema, que hemos visto en la endecatupla bungeana (o en la **Figura 1**), a la par que pone el foco en las personas, en una hermenéutica crítica (Cortina, 1996), alejada de una concepción meramente instrumental de la tecnología (Bunge, 2019 [1974]; Verbeek, 2005, 2011). De este modo, en el caso de la IA, una posible propuesta inspirada en Pareto *et al.* (2021) podría ser:

- *Bienestar:* incluye cuestiones como la privacidad y el control de datos, donde el uso de IA implica riesgos de vigilancia masiva y vulneración del consentimiento informado (Véliz, 2021). La autonomía humana también se ve amenazada, ya que una dependencia excesiva de la IA podría reducir la capacidad de toma de decisiones de los usuarios y ahondar en las problemáticas derivadas de la descarga cognitiva (causando un empeoramiento del aprendizaje, por ejemplo). Asimismo, la IA puede influir en la percepción de la identidad y dignidad humanas, propagando sesgos y estereotipos. Otros aspectos fundamentales en el bienestar incluyen la seguridad (tanto física como psicológica), el aislamiento y la pérdida del contacto humano y el impacto en las habilidades psicológicas y morales humanas.

- *Cuidado:* se centra en la legitimidad de la introducción de la IA en ámbitos tradicional e inherentemente humanos, como la educación, la salud o la asistencia social. La calidad de la práctica y la confianza en la IA son factores críticos, ya que su uso puede alterar las relaciones profesionales y la organización de roles. La robótica asistencial (Pareto *et al.*, 2021) es un buen ejemplo. También se podría cuestionar si la IA redefine el concepto mismo de cuidado, cambiando la relación entre quienes proveen y reciben atención, entre quienes se podrán permitir en un futuro un cuidador humano, y los que quedarán al cuidado de las máquinas. Por ejemplo, ya están proliferando aplicaciones basadas en IA de soporte psicológico que algunos pretenden que reemplacen a psicólogos profesionales. Los principios de la Declaración de Barcelona en este ámbito deberían estar muy presentes, empezando por la prudencia. Los riesgos son muchos; uno mencionado es que se acabe confiando más en la respuesta técnica que en la humana.

- *Justicia:* abarca la justicia distributiva. Es decir, la equidad en el acceso a los beneficios y costos de la IA, incluyendo su impacto en el mercado laboral y en la inclusión social. La justicia debe ser entendida más allá de los derechos individuales, como justicia social y global, de forma que no se ignoren los impactos de la IA en otras partes del mundo (**Figura 1**). La política de la tecnología relacionada con la IA es relevante, ya que sus objetivos y valores subyacentes pueden estar influenciados por intereses comerciales y gubernamentales. Además, la toma de decisiones automatizada plantea problemas de responsabilidad y transparencia, especialmente en sectores como la justicia o la sanidad. La automatización tiene sus ventajas, pero también genera conflictos, tanto técnicos como filosóficos y éticos (Innerarity, 2025). Finalmente, la sostenibilidad ecológica de la IA es un aspecto emergente, dado su elevado consumo de energía, el extractivismo y los residuos electrónicos que genera (Crawford, 2021).

Todos estos aspectos, por ejemplo, los ha intentado incorporar, con sus luces y sombras, la regulación de la IA de la Unión Europea (Comisión Europea, 2024). En la **Tabla 1** se han intentado agrupar y sintetizar algunos problemas éticos relacionados con la IA, según estas tres dimensiones. Se trataría de problemas abordables tanto desde la investigación filosófica como desde la educación, como se verá posteriormente.

**Tabla 1. Síntesis de algunos riesgos y problemas éticos relacionados con la IA, según las dimensiones de bienestar, cuidado y justicia**

| Dimensión | Problema ético | Descripción |
|---|---|---|
| Bienestar | Privacidad y control de los datos | Riesgo de vigilancia, falta de transparencia en el manejo de datos, con potenciales sesgos. |
| | Manipulación y engaño | IA generando interacciones engañosas, creando bulos y vínculos no auténticos. |
| | Autonomía humana | Dependencia de la IA, reducción de habilidades cognitivas o en la toma de decisiones. |
| | Pérdida de contacto humano | Sustitución de interacciones humanas por interacciones automatizadas con IA. |
| | Seguridad | Riesgos de accidentes, amenazas psicológicas, falta de confiabilidad. |
| | Dignidad e identidad | Riesgo de objetificación de las personas, con impacto en la autoestima, la percepción de autoeficacia y la autoimagen. |
| Cuidado | Legitimidad de la IA en prácticas humanas | Cuestionamiento y limitación de su uso en educación, salud y asistencia. |
| | Calidad y sesgos sociales de la práctica | Reducción de la calidad del servicio debido a automatización excesiva, pérdida de contacto humano y sesgos socioeconómicos. |
| | Confianza (o temor) a la automatización | Respuesta emocional y dependencia de sistemas de IA que pueden fallar y ser opacos, sesgados e incluso erráticos. |
| | Redefinición del cuidado | Cambios en el significado y las prácticas del cuidado humano, y redefinición del cuidado artificial y de su marco legal. |
| Justicia | Justicia distributiva | Desigualdades en el acceso y distribución de los beneficios de la IA. Se democratizan y distribuyen los costes, pero no los beneficios económicos. |
| | Política de la tecnología IA | Intereses comerciales y gubernamentales influyendo en la regulación, y en desarrollos con intereses particulares, en detrimento del interés social. |
| | Responsabilidad y transparencia | Dificultades para asignar responsabilidades en decisiones algorítmicas, y falta de transparencia (y de explicabilidad) en el entrenamiento y en los resultados. |
| | Sostenibilidad ecológica | Consumo de energía, agua y generación de residuos electrónicos. |
| | Impacto social global | Explotación de personas en países que no velan por los derechos laborales, donde las multinacionales delegan tareas mal pagadas, como el etiquetado de datos. |

Fuente: elaboración propia, siguiendo la propuesta de Pareto *et al.* (2021).

Estos marcos generales, por supuesto, tienen muchas aristas que requieren de una investigación filosófica y científica profunda, y deben concretarse según las áreas en las que la IA impacte. En el caso de la educación, como propuesta práctica, sería muy interesante abordar la formación de los estudiantes en la enseñanza obligatoria siguiendo las problemáticas de la **Tabla 1**, y por supuesto vinculada a los currículos educativos oficiales. En el caso de la secundaria, por ejemplo, hay algunos precedentes de propuestas concretas de tecnoética, en el marco curricular de la enseñanza en Cataluña (Ramírez Marqués, 2020; Martínez, 2021; Muñoz Cuenca, 2022), y, si se me permite, previos a la burbuja de la IA de finales de 2022.

Posteriormente, tras la irrupción especialmente de los Grandes Modelos de Lenguaje (*large language models* o LLM, en sus siglas en inglés) de la IA generativa, a tal efecto, la UNESCO propuso un marco competencial para docentes (Cukurova & Miao, 2024) y otro para estudiantes (Miao & Shiohira, 2024) que incluía (por fin) la dimensión ética dentro de una matriz bidimensional con cinco aspectos que evolucionan a través de tres niveles de progresión: adquisición, profundización y creación. Para abordar en la educación estos tres niveles de progresión en la enseñanza de la ética de la IA, es fundamental estructurar el aprendizaje en cada dimensión, adaptándose, por supuesto a los currículos oficiales que, no obstante, suelen ir por detrás de la innovación tecnológica. A continuación, se describe sucintamente cómo implementar estos niveles propuestos por la UNESCO en cada aspecto, ampliando el enfoque general de Hernández-Fernández (2024b):



*1. Mentalidad centrada en las personas:*

- Adquirir: enseñar en el aula la importancia de la agencia humana en la interacción con la IA, destacando que las personas deben tener control sobre las decisiones asistidas por la tecnología.
- Profundizar: estudiar casos donde la IA influye en la toma de decisiones y fomentar la reflexión sobre la responsabilidad individual en su uso, así como la responsabilidad social y de los políticos y administradores.
- Crear: diseñar proyectos donde los estudiantes evalúen y propongan formas donde la IA pueda fortalecer la responsabilidad social y la equidad, de forma crítica.

*2. Ética de la IA:*

- Adquirir: introducir los principios éticos fundamentales de la IA, como los de la Declaración de Barcelona, fomentando valores como la transparencia, la equidad o la privacidad.
- Profundizar: explorar dilemas éticos en el uso de IA, promoviendo un debate crítico en el aula sobre sus implicaciones en diferentes contextos, como la integración de algoritmos de decisión en los coches autónomos (véase, por ejemplo: Ramírez Marqués, 2020).

- Crear: desarrollar normas o directrices éticas concretas, en colaboración con otros estudiantes y docentes, para regular el uso de IA en distintas áreas de su actividad académica y cotidiana.

*3. Fundamentos y aplicaciones de la IA:*

- Adquirir: explicar los conceptos básicos de la IA, sus definiciones básicas, así como fundamentos sobre el funcionamiento del aprendizaje automático y los algoritmos.
- Profundizar: aplicar técnicas de IA a problemas concretos en distintas disciplinas, en especial en las materias de tecnología de secundaria.
- Crear: desarrollar soluciones innovadoras con IA, integrando herramientas avanzadas en proyectos propios, adecuados al nivel del curso.

*4. Pedagogía de la IA:*

- Adquirir: utilizar herramientas de IA para la enseñanza asistida, como sistemas de tutoría inteligente, que sean transparentes y de código abierto.
- Profundizar: integrar la IA en estrategias pedagógicas, adaptando la enseñanza a las necesidades de los estudiantes, atendiendo a la diversidad.
- Crear: diseñar nuevas metodologías educativas potenciadas por la IA, explorando su impacto real en el aprendizaje, sin olvidar los aspectos éticos, y descartando el uso de la IA cuando empeore el aprendizaje (luchando así contra el falso mito de la autonomía).



*5. IA para el desarrollo personal y profesional:*

- Adquirir: comprender cómo la IA puede facilitar el aprendizaje a lo largo de la vida, y ayudar a los estudiantes que no posean capacidad económica para pagar un profesor de refuerzo, por ejemplo.
- Profundizar: usar la IA para mejorar procesos organizacionales y la formación dentro de instituciones educativas.
- Crear: aplicar la IA para potenciar el desarrollo profesional, promoviendo cambios en las metodologías, la estructura del trabajo y el aprendizaje.

De este modo, un enfoque progresivo y amplio permitiría que la enseñanza de la ética de la IA evolucione desde la adquisición y comprensión básica de sus principios (primer nivel), la profundización en problemáticas concretas como las de la **Tabla 1** (segundo nivel), hasta la implementación activa y la innovación responsable (tercer nivel), siempre adecuándose al nivel y curso de los alumnos (Cukurova & Miao, 2024; Miao & Shiohira, 2024). En este sentido, queda como reto de la didáctica de la tecnología, o de las diversas disciplinas curriculares, en cuanto la tecnología se encuentra imbricada inevitablemente en todas las asignaturas, el proponer actividades y contenidos relacionados con las tres dimensiones de bienestar, cuidado y justicia (Pareto *et al.*, 2021).

## Conclusiones

Cincuenta años después de la conferencia de Haifa, en la que Bunge acuñó el término "tecnoética", en contraposición tácita a la bioética emergente, no podemos aquí sino recuperar el decálogo de conclusiones con las que Bunge (2019 [1974]) sentenciaba el recorrido esperable a aquella, entonces, disciplina emergente:

> "(i) A diferencia de la ciencia básica o pura, que es intrínsecamente valiosa o, en el peor de los casos, carece de valor, la tecnología puede ser valiosa o perjudicial según los fines que sirva. Por lo tanto, es necesario someter la tecnología a controles morales y sociales. (ii) La tecnología perversa solo puede eliminarse descartando sus fines dañinos. Y los malos usos de la buena tecnología no se corrigen ni se evitan frenando la investigación tecnológica, sino fomentando una tecnología aún mejor y haciéndola moral y socialmente sensible. (iii) El tecnólogo, como cualquier otra persona, es personalmente responsable de todo lo que diseña, planifica, recomienda o ejecuta. Por lo tanto, es tan digno de elogio o condena como cualquier otro. O incluso más, dado el carácter racional y deliberado de sus decisiones. (iv) El tecnólogo es responsable no solo ante su empleador y sus colegas profesionales, sino también ante todos los que puedan verse afectados por su trabajo. Su preocupación fundamental debería ser el bienestar público. (v) El tecnólogo que contribuye a aliviar los problemas sociales o a mejorar la calidad de vida es un benefactor público. Pero aquel que, a sabiendas, contribuye a deteriorar la calidad de vida, engaña al público diseñando productos inútiles o nocivos, o difunde información falsa, es un criminal. (vi) Dado que el especialista no puede abordar todos los problemas multilaterales y complejos que plantean los proyectos tecnológicos a gran escala, estos deberían encomendarse a equipos de expertos en diversos campos, incluidos los científicos sociales aplicados, y someterse al escrutinio y control públicos. (vii) Considerando el fuerte impacto de la tecnología en la sociedad y el medio ambiente, el tecnólogo debería compartir el poder con el administrador y el político. La tecnocracia global, o el gobierno de expertos en todos los campos de la acción humana, no es una amenaza, sino una promesa, especialmente si sus decisiones se toman con la aprobación del público. (viii) Los tecnólogos deberían abordar sus propios problemas morales, en lugar de fingir que estos pueden transferirse a los administradores o políticos. (ix) Los tecnólogos deberían contribuir a modernizar la ética, intentando construir una tecnoética como la ciencia de la conducta correcta y eficiente de ellos mismos. (x) La ética no saldrá de su estancamiento actual si los filósofos persisten en ignorar la experiencia moral de los

> tecnólogos, si no prestan atención a cómo la tecnología fundamenta sus prescripciones o normas, y si se niegan a emplear los recursos formales (lógicos y matemáticos) que brindan claridad y precisión" (Bunge, 2019 [1974], pp. 135-136).

En el momento actual, es más necesario que nunca un análisis de este decálogo de Bunge (2019 [1974]) sobre la tecnoética, en relación con los problemas actuales de la IA, especialmente la IA generativa. En general, podría concluirse:

i. La IA, como toda tecnología, debe someterse a controles morales y sociales. Las IA generativas pueden amplificar la desinformación, generar noticias falsas o manipular la opinión pública mediante bulos profundos (*deepfakes*). Es esencial establecer regulaciones y controles éticos que limiten su uso indebido y garanticen su alineación con el bien común (Comisión Europea, 2024).
ii. La IA perversa solo se elimina descartando sus fines dañinos. No basta con prohibir tecnologías problemáticas; hay que evitar y perseguir sus usos nocivos. Un ejemplo es el abuso de IA en la creación de contenido ilícito o falso, que no se resuelve eliminando la IA en sí, sino desarrollando herramientas para detectar y frenar la propagación de desinformación, o prohibiendo aplicaciones sin ningún uso bueno (Hernández-Fernández, 2024, 2025c).
iii. Los tecnólogos son responsables de lo que diseñan, planean y ejecutan. Los desarrolladores de IA deben asumir la responsabilidad de los sesgos en sus modelos, del impacto de la automatización en el empleo y del uso de la IA en vigilancia masiva (Véliz, 2020). No pueden excusarse en una supuesta neutralidad de la tecnología (Verbeek, 2011), máxime si sus creaciones amplifican desigualdades o vulneran derechos de las personas.
iv. Los tecnólogos son responsables ante la sociedad en su conjunto. La IA afecta a millones de personas, desde su uso en redes sociales, hasta en la toma de decisiones bancarias (si se concede o no un crédito), educativas (en las calificaciones académicas) o médicas (en diagnósticos médicos). Los profesionales y tecnólogos no pueden actuar solo en interés de sus empleadores o accionistas; deben considerar el impacto social de sus creaciones y rendir cuentas (Pareto *et al.*, 2021).
v. La IA puede ser beneficiosa o perjudicial según su uso. Sin abundar en el mito de la neutralidad, no debemos negar que la IA generativa puede ayudar en la educación, la creatividad o la medicina, pero también se usa para manipular elecciones, generar bulos o facilitar el fraude. Es clave educar en un uso crítico, responsable y ético que maximice los beneficios y minimice los daños de la IA (Hernández-Fernández, 2024), como de toda tecnología, sin olvidar su regulación (Coeckelbergh, 2019).
vi. La complejidad de la IA exige equipos multidisciplinares y control público. El desarrollo de IA no debe quedar solo en manos de ingenieros, tecnólogos y empresarios. Es necesario incorporar filósofos, juristas, sociólogos y expertos en

derechos humanos para evaluar impactos y garantizar su regulación democrática, como se ha visto en el caso de la robótica asistencial (Pareto & Torras, 2024).

vii. La tecnocracia global no es una amenaza si sus decisiones se someten a la aprobación pública. El poder de la IA en la toma de decisiones (desde créditos bancarios o diagnósticos clínicos, hasta la seguridad nacional) y en el control masivo de datos no puede quedar en manos de unas pocas corporaciones, oligopolios o multinacionales tecnológicas. Se necesita supervisión pública y mecanismos democráticos para evitar que la IA se convierta en un instrumento de control social en el capitalismo de vigilancia (Véliz, 2020; O'Neil, 2018).

viii. Los tecnólogos deben abordar los problemas morales de la IA. El sesgo algorítmico, la explotación laboral en el entrenamiento de los sistemas de IA, o la privacidad de los usuarios, no pueden ser problemas delegados a reguladores externos. Los creadores de IA deben integrar la ética en el diseño desde el principio y contemplar todas las fases del diseño tecnológico (Hernández-Fernández, 2025).

ix. Es necesaria una tecnoética que guíe la evolución de la IA. Más allá de regulaciones reactivas, se necesita una ética específica para la IA que combine principios filosóficos con conocimientos técnicos, ayudando a diseñar IA seguras, transparentes, explicables y alineadas con los valores humanos (Comisión Europea, 2024).

x. La tecnoética de la IA no debe ignorar la experiencia de los tecnólogos. Los debates éticos sobre la IA tampoco pueden quedar solo en manos de filósofos alejados de la práctica tecnológica. Son cruciales los equipos transdisciplinares y que los propios desarrolladores participen en la construcción de normas y estándares éticos que guíen la evolución de la IA, máxime en aras de la reducción del impacto ambiental y la sostenibilidad (Mokiy, 2023; Crawford, 2021).



En conclusión, cincuenta años después de la conferencia de Haifa, transmitir el espíritu de la tecnoética de Bunge es fundamental, cuando no urgente. Me sorprende la ausencia de sus reflexiones y postulados en buena parte de los estudios críticos recientes sobre la IA. La irrupción de la IA generativa plantea desafíos complejos e inéditos en la intersección entre tecnología, sociedad y ética, desde la proliferación de desinformación hasta el riesgo de vigilancia masiva y sesgos algorítmicos. Retomar el enfoque progresivo y amplio que Bunge propuso implica no solo regular la IA para minimizar sus riesgos, sino también integrarla en la enseñanza y el debate público, garantizando su desarrollo bajo principios de bienestar, cuidado y justicia (Pareto *et al.*, 2021). En este sentido, la tecnoética no debe quedarse en una reflexión abstracta, sino traducirse en una praxis crítica y necesariamente innovadora, en la educación como en otros ámbitos, que guíe la evolución de la IA y sus usos hacia un horizonte de responsabilidad social y sostenibilidad. En caso contrario, las consecuencias pueden ser nefastas para la humanidad.

**Financiamiento**

Esta publicación es parte del proyecto PID2024-155946NB-I00 del Ministerio de Ciencia, Innovación y Universidades (MICIU), la Agencia Estatal de Investigación (AEI/10.13039/501100011033) y del European Social Fund Plus (ESF+), HORIZON-101234399-BSC-AI-Factory, así como del reconocimiento de la Generalitat de Catalunya 2021 SGR 01266.

**Reconocimiento de uso de IA**

El autor reconoce el uso de los modelos de lenguaje integrados en HuggingChat para la revisión de la traducción del resumen y las palabras clave al inglés y al portugués, así como también para dar formato en la generación de la **Tabla 1**. Más allá de este reconocimiento, se declara como único autor del artículo y es completamente responsable del contenido incluido en él.

**Agradecimientos**

Debo agradecer a Mario Bunge su mentoría intelectual y la correspondencia que establecimos en sus últimos años, tanto en lo referente a la tecnoética como a la filosofía en general. También a su alumno aventajado, Alfons Barceló, por los múltiples debates que en parte se han reflejado aquí.



**Bibliografía**

Altmann, G., & Koch, W. A. (2011). Systems: New paradigms for the human sciences. Berlín: Walter de Gruyter.

Boden, M. A. (2010). Creativity and art: Three roads to surprise. Oxford: Oxford University Press.

Bunge, M. (2019 [1974]). Por una tecnoética. Conferencia «Ethics in an Age of Pervasive Technology», Technion, Haifa. En Filosofía de la tecnología (131-143). Barcelona: Edicions IEC-Iniciativa Digital Politècnica.

Bunge, M. (2012). Filosofía de la tecnología y otros ensayos. Lima: Fondo Editorial.

Bunge, M. (2013). Pseudociencia e ideología. Pamplona: Laetoli.

Bunge, M. (2019). Filosofía de la tecnología. Barcelona: Edicions IEC-Iniciativa Digital Politècnica. Recuperado de: http://hdl.handle.net/2117/169030.

Carbonell, E. & Sala, R. (2002). Aún no somos humanos: propuestas de humanización para el tercer milenio. Barcelona: Península.

Coeckelbergh, M. (2019). Artificial intelligence: Some ethical issues and regulatory challenges. Technology and regulation, 2019, 31-34. DOI: https://doi.org/10.71265/a9yxhg88.

Comisión Europea (2024). Ley de la IA, Reglamento (UE) 2024/1689 por el que se establecen normas armonizadas sobre inteligencia artificial. Recuperado de: https://digital-strategy.ec.europa.eu/es/policies/regulatory-framework-ai.

Cortina, A. (1996). El estatuto de la ética aplicada. Hermenéutica crítica de las actividades humanas. Isegoría, 13, 119–127.

Crawford, K. (2021). Atlas of AI: power, politics, and the planetary costs of artificial intelligence. New Haven: Yale University Press.

Cukurova, M. & Miao, F. (2024). AI competency framework for teachers. UNESCO Publishing. Recuperado de: https://unesdoc.unesco.org/ark:/48223/pf0000391104.



Diéguez, A. (2024). Pensar la tecnología. Barcelona: Shackleton books.

Du Sautoy, M. (2020). Programados para crear. Barcelona: Acantilado.

Estampa (2022). Cartografía de la IA. Proyecto Artístico disponible en abierto en: https://cartography-of-generative-ai.net/.

EU-HLEG (2019). A definition of AI: main capabilities and disciplines. Bruselas: Comisión Europea. Recuperado de: https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines .

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. Information Systems Research, 33(2), 678-696. DOI: https://doi.org/10.1287/isre.2021.1079

Gordon, J. S. & Nyholm, S. (2024). Ethics of Artificial Intelligence. En The Internet Encyclopedia of Philosophy. Recuperado de: https://iep.utm.edu/ethics-of-artificial-intelligence/

Hernández-Fernández, A. (2024a). Tecnoética en el uso de la inteligencia artificial: el derecho a no ser desnudado. En: Perspectivas contemporáneas en educación:

innovación, investigación y transformación (1234-1251). Dykinson. Recuperado de: http://hdl.handle.net/2117/422475.

Hernández-Fernández, A. (2024b). Inteligencia artificial para docentes según la UNESCO. Educational Evidence. Recuperado de: https://educationalevidence.com/inteligencia-artificial-para-docentes-segun-la-unesco/.

Hernández Fernández, A. (2025a). Inteligencia artificial y diseño. Revista Gráfica, 13(26), 265-288. DOI: https://doi.org/10.5565/rev/grafica.461.

Hernández Fernández, A. (2025b). Intel·ligència artificial i creativitat en l'ensenyament artístic. Revista Imatges Colbacat, 5, 11-20. Recuperado de: https://hdl.handle.net/2117/430178.

Hernández Fernández, A. (2025c). Desnudando mujeres: sesgo de género en la inteligencia artificial generativa de imagen. En: WSCITECH25 (12-24). Recuperado de: https://hdl.handle.net/2117/428809.

Holmes, W., Bialik, M. & Fadel, C. (2019). Artificial intelligence in education: promises and implications for teaching and learning. Center for Curriculum Redesign.

Innerarity, D. (2025). Una teoría crítica de la inteligencia artificial. Barcelona: Galaxia Gutenberg.

Kahneman, D., Sibony, O. & Sunstein, C. R. (2021). Ruido: Un fallo en el juicio humano. Barcelona: Debate.

López de Mántaras, R. (2023). 100 coses que cal saber sobre intel·ligència artificial. Valls: Cossetània Edicions.

Ramírez Marqués, E. (2020). Tecnoètica a l'Educació Secundària Obligatòria. Recuperado de: http://hdl.handle.net/2117/333380.

Maleki, N., Padmanabhan, B. & Dutta, K. (2024). AI hallucinations: a misnomer worth clarifying. En 2024 IEEE conference on artificial intelligence (CAI) (133-138). IEEE. DOI: https://doi.org/10.1109/CAI59869.2024.00033.

Martínez, N. (2021). Tecnoètica al segon cicle de l'Educació Secundària Obligatòria (3r i 4t d'ESO). Recuperado de: http://hdl.handle.net/2117/349488.

Miao, F., & Shiohira, K. (2024). AI competency framework for students. UNESCO Publishing. Recuperado de: https://unesdoc.unesco.org/ark:/48223/pf0000391105.

Mokiy, V. (2023). Transdisciplinarity and Artificial Intelligence in the Service of Sustainable Development of Society. Artificial Intelligence and Human Mediation, 61.

Recuperado de: Artificial-Intelligence-and-Human-Mediation-Invited-editors-and-co-authors-Leonardo-Costa-and-Mariana-Thieriot.pdf.

Muñoz Cuenca, A. (2022). Futuro, ética y digitalización: una propuesta de introducción de la tecnoética en la asignatura "Tecnología e Ingeniería". Recuperado de: http://hdl.handle.net/2117/370731.

O'neil, C. (2018). Armas de destrucción matemática: cómo el big data aumenta la desigualdad y amenaza la democracia. Madrid: Capitán Swing Libros.

Ortega y Gasset, J. (1982 [1939]). Meditación de la técnica y otros ensayos sobre ciencia y filosofía. Buenos Aires: Espasa-Calpe Argentina. En: Obras, vol. 21. Madrid: Revista de Occidente.

Pareto, J. & Torras, C. (2024). To each Technology Its Own Ethics? A Reply to Sætra & Danaher. Philosophy & Technology, 37(3), 107. DOI: https://doi.org/10.1007/s13347-024-00798-w.

Pareto, J., Román, B. & Torras, C. (2021). The ethical issues of social assistive robotics: A critical literature review. Technology in Society, 67, 101726. DOI: https://doi.org/10.1016/j.techsoc.2021.101726.



Quintanilla, M. Á. (2017). Tecnología: un enfoque filosófico y otros ensayos de filosofía de la tecnología. Ciudad de México: Fondo de Cultura Económica.

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. AJOB neuroscience, 11(2), 88-95. DOI: https://doi.org/10.1080/21507740.2020.1740350.

van der Rijt, J. W., Mollo, D. C. & Vaassen, B. (2025). AI Mimicry and Human Dignity: Chatbot Use as a Violation of Self-Respect. arXiv preprint arXiv:2503.05723.

Véliz, C. (2021). Privacidad es poder. Madrid: Debate.

Verbeek, P. P. (2005). What things do: Philosophical reflections on technology, agency, and design. University Park: Pennsylvania State Univ. Press.

Verbeek, P. P. (2011). Moralizing technology: Understanding and designing the morality of things. Chicago: University of Chicago press.

Wingström, Roosa, Hautala, Johanna & Lundman, Riina (2024). Redefining creativity in the era of AI? Perspectives of computer scientists and new media artists. Creativity Research Journal, 36(2), pp. 177-193. DOI: https://doi.org/10.1080/10400419.2022.2107850.

# Automatizar el mundo, gobernar la vida. Debates contemporáneos sobre inteligencia artificial y poder desde una perspectiva ética y latinoamericana

# Automatizar o mundo, governar a vida. Debates contemporâneos sobre inteligência artificial e poder a partir de uma perspectiva ética e latino-americana

# *Automating the world, Governing Life. Contemporary Debates on Artificial Intelligence and Power from an Ethical and Latin-American Perspective*

**Hernán Gabriel Borisonik** *

Este artículo aborda los principales problemas ético-políticos vinculados con el desarrollo y la implementación de sistemas de inteligencia artificial (IA), situándolos en el marco más amplio de la digitalización de la experiencia vital y sus efectos en la producción de subjetividades contemporáneas. A partir de una perspectiva de filosofía política crítica, se examinan las tensiones entre las promesas tecnológicas de la IA y sus efectos concretos en contextos sociales desiguales. El artículo se organiza en tres partes. En primer lugar, se discuten los principales enfoques teóricos recientes sobre la IA desde perspectivas tecnocríticas, tecnooptimistas, marxistas y poshumanistas. En segundo lugar, se analizan los problemas específicos que surgen de la implementación de estos sistemas, especialmente el sesgo algorítmico, las nuevas formas de vigilancia y extracción de datos, y los efectos sobre la subjetivación. Por último, se presentan los debates actuales en torno a la gobernanza de la IA, destacando tanto las agendas globales como las discusiones emergentes en América Latina. A modo de conclusión, se propone superar las posturas tecnofóbicas y solucionistas, y se recupera la idea de "tecnodiversidad" (Hui, 2021) como un horizonte posible para pensar desarrollos tecnológicos situados, heterogéneos y atentos a las especificidades regionales.

**Palabras clave:** inteligencia artificial; ética; economía política; subjetividad; digitalidad



* Doctor en ciencias sociales por la Universidad de Buenos Aires (UBA), Argentina. Profesor adjunto regular en la Escuela de Humanidades de la Universidad Nacional de San Martín (UNSAM), Argentina. Investigador adjunto del CONICET y del Laboratorio de Investigación en Ciencias Humanas (LICH-UNSAM). Director del Centro Ciencia y Pensamiento (CCyP-UNSAM). *Visiting scholar* de la Universitat Oberta de Catalunya (UOC). Publicó, entre otros, los libros *Dinero sagrado*, *Soporte* y *Persistencia de la pregunta por el arte*. Correo electrónico: hborisonik@unsam.edu.ar. ORCID: https://orcid.org/0000-0003-3247-043X.

Este artigo aborda os principais problemas ético-políticos associados ao desenvolvimento e à implementação de sistemas de inteligência artificial (IA), situando-os dentro da estrutura mais ampla da digitalização da experiência de vida e seus efeitos sobre a produção de subjetividades contemporâneas. Sob a perspectiva de uma filosofia política crítica, ele examina as tensões entre as promessas tecnológicas da IA e seus efeitos concretos em contextos sociais desiguais. O artigo está organizado em três partes. Primeiro, ele discute as principais abordagens teóricas recentes da IA a partir das perspectivas tecno-crítica, tecno-otimista, marxista e pós-humanista. Em seguida, são analisados os problemas específicos decorrentes da implementação desses sistemas, especialmente o viés algorítmico, as novas formas de vigilância e mineração de dados e os efeitos sobre a subjetivação. Por fim, são apresentados os debates atuais sobre a governança da IA destacando tanto as agendas globais quanto as discussões emergentes na América Latina. Nas conclusões, propomos a superação de posições tecnofóbicas e solucionistas e recuperamos a ideia de "tecno-diversidade" (Hui, 2021) como um horizonte possível para pensar em desenvolvimentos tecnológicos situados, heterogêneos e atentos às especificidades regionais.

**Palavras-chave:** inteligência artificial; ética; economia política; subjetividade; digitalidade

*This article addresses the main ethical-political problems associated with the development and implementation of artificial intelligence (AI) systems, situating them within the broader framework of the digitalization of life experience and its effects on the production of contemporary subjectivities. From a critical political philosophy perspective, it examines the tensions between the technological promises of AI and its concrete effects in unequal social contexts. The article is organized in three parts. Firstly, it discusses the main recent theoretical approaches to AI from techno-critical, techno-optimistic, Marxist and post-humanist perspectives. Secondly, the specific problems arising from the implementation of these systems are analyzed, especially algorithmic bias, new forms of surveillance and data mining, and effects on subjectification. Finally, current debates around AI governance are presented, highlighting both global agendas and emerging discussions in Latin America. In the conclusions, we propose overcoming technophobic and solutionist positions, and recovering the idea of "techno-diversity" (Hui, 2021) as a possible horizon for thinking about technological developments that are situated, heterogeneous and attentive to regional specificities.*

***Keywords:*** *artificial intelligence; ethics; political economy; subjectivity; digitality*

**Introducción: máquinas que piensan, ¿sociedades que obedecen?**

En las últimas décadas, la inteligencia artificial (IA) ha pasado de ser una especulación científica a convertirse en un actor central en la reorganización de la experiencia humana. Su influencia se extiende desde la automatización del trabajo hasta la reconfiguración de la comunicación, el conocimiento y la toma de decisiones. A través de sistemas de aprendizaje automático, procesamiento de datos y generación de contenidos, la IA participa en la modulación de los vínculos sociales y en la producción de subjetividades, marcando una nueva etapa en la digitalización de la experiencia vital (Rouvroy & Berns, 2013; Floridi *et al.*, 2018). Esta digitalización no es solo un proceso técnico, sino una transformación estructural que redefine el acceso al mundo, mediando las relaciones entre individuos, instituciones y estructuras de poder. Y si bien en los últimos años han aparecido varios trabajos acerca de la disrupción tecnológica (Torres, 2023; García Serrano, 2016; Spitzer, 2023; Krauss, 2023) y las formas novedosas de interacción que acompaña (y promueve) este invento (Lupton, 2016), menos se ha avanzado, en este contexto, en los dilemas éticos fundamentales que vienen de la mano de la IA. ¿Cómo garantizar que los sistemas algorítmicos operen de manera justa? ¿Es posible asignar responsabilidad moral a una máquina? ¿En qué medida la IA reproduce o amplifica desigualdades estructurales? Estas preguntas reflejan la profundidad del desafío: no se trata únicamente de regular tecnologías emergentes, sino de interrogar cómo estas tecnologías reconfiguran la propia noción de agencia, alteran las condiciones de posibilidad del conocimiento y desplazan los marcos tradicionales de deliberación moral (UNESCO, 2021).



El debate ético en torno a la IA se inscribe en una discusión más amplia sobre la relación entre técnica, subjetividad y poder en la era digital. Diferentes corrientes filosóficas han abordado esta problemática desde enfoques normativos clásicos, como la deontología kantiana y el utilitarismo, hasta perspectivas más situadas, como la ética del cuidado y las teorías críticas de la tecnología. Sin embargo, estos marcos presentan limitaciones cuando se aplican a sistemas dinámicos y no humanos de toma de decisiones. ¿Hasta qué punto las categorías tradicionales de la ética son suficientes para comprender la acción algorítmica? ¿Es posible desarrollar una perspectiva ética que contemple la especificidad sociotécnica de la IA sin caer en determinismos tecnológicos o en discursos excesivamente abstractos? Este artículo contribuye a este debate mediante una revisión crítica de los principales enfoques éticos sobre la IA, explorando sus alcances y limitaciones. Los problemas que sirven como piedra fundamental para las próximas páginas son: i) qué es la inteligencia artificial; ii) cuáles son sus principales aplicaciones; y iii) cuáles son los aspectos cardinales a tener en cuenta a la hora de buscar marcos éticos (y regulatorios). A partir de ellos, se esboza una propuesta de ética situada, que contemple los contextos específicos en los que operan estos sistemas y la manera en que reconfiguran las prácticas sociales y los regímenes de verdad. Se argumenta que una aproximación contextualizada permite abordar de manera más precisa los dilemas emergentes y evita respuestas normativas que, al abstraerse de las condiciones materiales e institucionales de la IA, resultan insuficientes para enfrentar sus implicaciones reales.

En vistas de lo anterior, la estructura del artículo se organiza en cuatro secciones principales, cada una de las cuales aborda un aspecto específico de la relación entre IA y ética desde una perspectiva vinculada con la filosofía política, con especial atención a los debates contemporáneos y las disputas teóricas más recientes. En primer lugar, esta introducción tiene como objetivo presentar el problema general que orienta la investigación: la necesidad de descentrar las narrativas tecnofílicas y tecnofóbicas predominantes y analizar la IA como un objeto político y cultural, inserto en relaciones de poder, en tramas materiales concretas y en imaginarios geopolíticos diferenciados. En este sentido, se plantea que pensar la IA exige escapar tanto del solucionismo tecnológico como del fatalismo distópico, abriendo el análisis a las condiciones sociales, económicas y culturales que modelan su desarrollo y sus efectos.

La segunda parte constituye el marco teórico del trabajo y se propone como un recorrido crítico por las principales perspectivas contemporáneas en torno a la IA. Este recorrido no se organiza de modo exhaustivo, sino que selecciona algunos debates clave para la filosofía política y las ciencias sociales. En un primer momento, se recuperan los aportes de las tradiciones marxistas y posmarxistas, que analizan la IA como parte del desarrollo histórico del capitalismo y la automatización, enfatizando sus efectos sobre el trabajo, la extracción de valor y la organización social (Marx, 1990 [1867]; Dyer-Witheford, Kjøsen & Steinhoff, 2024). Luego, se abordan las posiciones tecnooptimistas y transhumanistas, que celebran el avance de la IA como superación de los límites humanos, pero que suelen reproducir lógicas profundamente desiguales y extractivistas (Kurzweil, 2005). En tercer lugar, se examinan las perspectivas centradas en la gubernamentalidad algorítmica y la extracción de datos, que conceptualizan la IA como un dispositivo de captura de la vida social y de modelización predictiva de las conductas (Zuboff, 2019; Rouvroy & Berns, 2013). Finalmente, se presentan las críticas desde el neorrealismo y los estudios de tecnología, que problematizan la dimensión geopolítica de la IA, su inserción en dinámicas de soberanía técnica y la necesidad de pensar alternativas situadas y pluralistas (Hui, 2021; Mejias & Couldry, 2019).



La tercera sección del artículo aborda el estado del arte, es decir, los principales abordajes empíricos y conceptuales que en los últimos años han problematizado las relaciones entre IA, trabajo, subjetividad y poder. En primer lugar, se analiza la literatura centrada en las transformaciones del trabajo y la automatización, en particular los estudios sobre plataformas digitales, microtrabajo y economía de datos (Gray & Suri, 2019). Luego, se consideran las investigaciones sobre IA y colonialismo de datos, que plantean la expansión algorítmica como una continuidad de dinámicas extractivas propias del capitalismo global (Crawford, 2021; Mohamed, Isaac & Png, 2020). Finalmente, se presenta un apartado específico centrado en América Latina, donde se recuperan trabajos que abordan las tensiones entre soberanía tecnológica, desarrollos locales de IA y reproducción de desigualdades regionales, a partir de estudios recientes que han comenzado a problematizar zonas específicas (como América Latina, o el feminismo) en este campo (Belli & Zingales, 2022; Álvarez Alvarez *et al*., 2021).

Por último, la cuarta parte se ocupa de presentar las conclusiones del trabajo. Allí se retoman los principales argumentos desarrollados y se propone un posicionamiento crítico frente a los discursos polarizados sobre la IA. Lejos de adherir a posturas plenamente pesimistas –que idealizan un pasado pretecnológico (o rechazan de plano cualquier avance)– o a soluciones plenamente automatizadas –que despolitizan las condiciones de producción de la tecnología–, se sostiene que uno de los caminos más promisorios consiste en recuperar y actualizar las ideas de tecnodiversidad propuestas por Hui (2021), pero repensándolas desde y para cada región (en nuestro caso, América Latina). Esta perspectiva es propuesta como la más fructífera para imaginar futuros tecnológicos no subordinados a las lógicas del capital global, sino anclados en saberes situados, en ecologías políticas propias y en la heterogeneidad cultural y técnica de la región.

## 1. La inteligencia artificial en el horizonte de la crítica: Silicon Valley y la gobernanza algorítmica

La irrupción de la inteligencia artificial como fuerza técnica, epistémica y económica ha desatado un vasto repertorio de discursos sobre sus alcances, riesgos y promesas. Como toda tecnología significativa, la IA no produce solamente transformaciones en el orden técnico, sino que también reconfigura las formas de vida y convivencia, los esquemas de inteligibilidad y, por supuesto, los regímenes de poder. En ese marco, es necesario pensar la IA no como un objeto aislado, sino como una mediación central para el siglo XXI, pero altamente densa en disputas políticas, cargada de historia, de ideología y de conflicto.



Diversos enfoques han surgido para abordar estos desafíos y cada uno de ellos ha aportado perspectivas únicas sobre cómo la IA debería integrarse en la sociedad. Desde la filosofía política y las ciencias sociales hasta la teoría crítica, la IA constituye un nuevo punto de condensación en el que convergen algunas de las tensiones fundamentales del capitalismo contemporáneo: la relación entre automatización y trabajo, la gubernamentalidad algorítmica, la mutación de las formas de subjetividad, el rol de los datos como nuevo recurso estratégico y la cuestión de la soberanía técnica. La deontología y el utilitarismo, como marcos éticos tradicionales, ofrecen enfoques distintos que podrían ser aplicados para pensar las especificidades de la IA. Así, mientras que el primero prioriza principios morales universales como la justicia, el segundo busca maximizar el bienestar general (Cortina, 2024; Moor, 1985). La ética del cuidado propone que la IA debe diseñarse con empatía y atención a las necesidades individuales, integrando valores humanos para fomentar el bienestar social (Danesi, 2025). Las teorías críticas de la tecnología denuncian cómo la IA puede reforzar las grandes desigualdades ya existentes y exigen un desarrollo más democrático, cuestionando las estructuras de poder detrás de su implementación (Xun, 2025). No faltan las miradas que resaltan a la IA como una herramienta que, en el contexto capitalista, intensifica la explotación laboral y consolida el dominio de la clase burguesa (Veraza Urtuzuástegui, 2024). La llamada ética de la virtud propone

que la IA debería diseñarse para promover virtudes humanas, como por ejemplo la prudencia y la justicia, fomentando el florecimiento moral de las personas (Schneider, 2019). A continuación, se propone un recorrido algo más profundo por algunas de las posturas que pueden ayudar a situar estas problemáticas.

Una primera línea crítica puede encontrarse en las tradiciones marxistas y posmarxistas, que han abordado la automatización como parte del desarrollo contradictorio del capital. Ya en el pensamiento del propio Marx (1990 [1867]), la maquinaria ha sido vista como un instrumento para la extracción de plusvalía relativa, al reducir el tiempo de trabajo necesario. En este marco, la automatización no libera del trabajo, sino que reestructura su explotación. Las tecnologías no son neutras, sino que encarnan lógicas de producción que las preexisten, que en este caso son sin duda alguna las del capital. De ahí que, en los hechos, la mayoría de los desarrollos en IA han sido afines a extender las formas del dominio sobre el tiempo, los recursos, los cuerpos y la cooperación social.

Autores como Dyer-Witheford, Kjøsen y Steinhoff (2024) han profundizado este enfoque, al analizar cómo la IA participa en una nueva fase del capitalismo cognitivo, donde los datos –extraídos, muchas veces voluntariamente, pero muchísimas más veces sin conocimiento de los sujetos, de la interacción social– se convierten en una fuente directa de valorización. Desde esta perspectiva, los algoritmos no son simplemente herramientas, sino formas de organización del trabajo abstracto. Lejos de la fantasía de la automatización plena, lo que se observa es una intensificación de la explotación: microtrabajo, *gig economy* (o “economía *gig*”, que refiere a la dependencia económica de la realización de trabajos temporales que a menudo se consiguen a través de plataformas digitales), sistemas de puntuación de desempeño y vigilancia extendida. Así, la IA deviene una forma de subsunción algorítmica del trabajo vivo.



Desde otro ángulo, contrario al anterior, existen enfoques acríticos (entre ellos, los transhumanistas) que tienden a considerar la IA como un vector de progreso inevitable y deseable. En esta narrativa, las capacidades de procesamiento, predicción y aprendizaje de los sistemas artificiales abrirían la puerta a una nueva era de eficiencia, racionalidad y superación de los límites humanos. Autores como Kurzweil (2005) proponen una visión en la que la convergencia entre IA, biotecnología y nanotecnología llevaría a la aparición de una “singularidad” tecnológica en la que la inteligencia artificial ya no solamente igualaría, sino que superaría a la humana en cada tarea en particular y en la coordinación de todas ellas en general. Estas posiciones suelen reproducir un imaginario tecnocientífico (las más de las veces arraigado en el desarrollo occidental) centrado en el dominio y el control, y minimizan los efectos negativos sociales, políticos y ecológicos que presenta la expansión tecnocapitalista. Su punto ciego no reside en el entusiasmo por la técnica, sino en la ceguera respecto a las condiciones históricas y materiales en las que esta se desarrolla. En el fondo, su promesa de emancipación se apoya en una concepción profundamente desigual del mundo: una élite tecnológicamente aumentada que es libre y cada vez más poderosa

frente a vastos sectores desposeídos. Para Brynjolfsson (2014), por ejemplo, la IA es un "motor de prosperidad" que, mediante ganancias de productividad, podría financiar rentas básicas universales. Sus estudios en MIT muestran cómo los algoritmos de logística aumentan un 30% la eficiencia en determinadas pymes. Pero mientras los optimistas promueven la autorregulación corporativa (Brynjolfsson asesora a Google), hay miradas que pueden observar el fenómeno menos interesadamente. Más moderado, Benkler (2006) propone modelos híbridos donde IA sirva a *commons* digitales, no solo a mercados.

En contraste, existen otros enfoques críticos que han puesto el acento en la dimensión epistémica y gubernamental de la IA. Zuboff (2019) ha denunciado el uso de la IA como herramienta de extracción de plusvalía conductual, en un entramado de plataformas que monetizan hasta los microgestos humanos bajo modelos de negocio basados en la vigilancia (Zuboff testificó en el caso *antitrust* contra Meta). Zuboff (2019) ha analizado cómo las plataformas digitales configuran un nuevo régimen de acumulación basado en la extracción de datos conductuales, que alimentan modelos predictivos orientados a la modulación de los comportamientos, en general con fines comerciales o electorales. En esta lógica, la IA no solo "conoce" el mundo, sino que lo reorganiza anticipadamente e incita determinadas prácticas y acciones. Desde esta óptica, gran parte del desarrollo tecnológico alrededor de la IA se trata de una forma de poder que opera mediante la captura de la experiencia y su conversión en patrones calculables, consolidando un régimen que Zuboff denomina "capitalismo de vigilancia". Este tipo de crítica resuena con los estudios foucaultianos acerca de la gubernamentalidad: la IA no impone, sino que orienta; no reprime, sino que prefigura. La política, en este escenario, se desplaza hacia la gestión de riesgos, la optimización de flujos y la administración de la contingencia. Como señalan Rouvroy y Berns (2013), asistimos a una nueva forma de "gubernamentalidad algorítmica" que desarticula el sujeto de derecho en favor de perfiles probabilísticos, sustituyendo la deliberación política por la preacción estadística. Y, de ese modo, la condición de ciudadanía se diluye bajo formas de sujeción ancladas en el mundo privado, como las plataformas virtuales, quedando sometida a los modos de circulación de los bienes digitales bajo los vectores direccionales que estos nuevos actores definen.



En este punto, un aporte clave para poder fragmentar o desglosar las miradas más abstractas o centradas en Europa y los Estados Unidos proviene de los estudios decoloniales y del Sur global, que han comenzado a pensar en la IA como parte de un proceso más amplio del colonialismo de datos. Autores como Ricaurte (2019) y Couldry y Mejías (2019) advierten que la extracción masiva de datos reproduce dinámicas de desposesión ya conocidas: no se apropian únicamente de los recursos naturales, sino también formas de vida, relaciones sociales y modos de saber. Desde esta óptica, la IA funciona como un dispositivo epistémico-colonial que traduce el mundo en términos funcionales al capital, borrando saberes, lenguas y prácticas no hegemónicas. La ética de la inteligencia artificial requiere un análisis que trascienda los marcos eurocéntricos dominantes. Como señalan D'Ignazio y Klein (2020) en su "feminismo de los datos", los sistemas de IA reproducen estructuras de poder

coloniales al invisibilizar saberes no occidentales y priorizar datos de poblaciones privilegiadas. Este enfoque decolonial revela cómo la IA, lejos de ser neutral, perpetúa lo que de Sousa Santos (2017) denominó "epistemicidio" (la destrucción de formas de conocimiento marginalizadas). A esto se suma la crítica de Morozov (2013) al "solucionismo tecnológico", esa fe ingenua en que algoritmos pueden resolver problemas sociales complejos sin cuestionar sus raíces estructurales. El caso de los sistemas de predicción policial en América Latina ejemplifica este reduccionismo: al entrenarse con datos de arrestos históricos (sesgados por racismo institucional), automatizan la criminalización de pobres y minorías (Ávila *et al.*, 2023).

Estas lecturas se complementan con perspectivas que provienen desde puntos de vista del feminismo y de la crítica racial, que han mostrado cómo los sistemas de IA tienden a amplificar los sesgos estructurales del presente. Investigadoras como Benjamin (2019) o Noble (2018) han documentado cómo la clasificación algorítmica perpetúa estereotipos, discrimina cuerpos racializados y refuerza la invisibilidad de colectivos históricamente marginados. En lugar de corregir injusticias, la IA muchas veces las sistematiza, reproduciendo exclusiones en clave técnica. Por su parte, la UNESCO (2021) ha propuesto la *Recomendación sobre la ética de la inteligencia artificial*, que busca traducir algunos principios éticos generalizables en acciones políticas concretas para garantizar que la IA beneficie a la humanidad, respetando los derechos humanos y promoviendo el desarrollo sostenible. Esta iniciativa representa un paso hacia la construcción de un marco ético y jurídico global que aborde las preocupaciones planteadas por las distintas perspectivas filosóficas.



Por otro lado, han emergido también voces que, sin perder el enfoque crítico, proponen reimaginar la técnica desde otros marcos culturales y ontológicos. Uno de los aportes más sugestivos en esta dirección es el de Hui (2021), quien propone una "cosmotécnica" que reconozca la diversidad de formas de pensar y producir tecnología. Su noción de "tecnodiversidad" plantea que el diseño técnico no puede reducirse a una racionalidad universal, sino que debe inscribirse en cosmologías particulares. Esta idea abre un espacio fértil para pensar una IA no occidental, no colonizante, no homogénea.

Las críticas más productivas, en tanto que posiciones posibles y en diálogo frente al mundo que viene, no son las que rechazan la IA en bloque, sino las que entienden que el problema no es la existencia de la tecnología, sino el tipo de relaciones sociales y políticas que la configuran y le dan sentido. La pregunta, entonces, no es solo qué hace la IA, sino quién hace la IA, es decir: quién la diseña, con qué fines, desde qué supuestos epistémicos y para qué formas de vida. Este enfoque no clausura el debate, sino que lo politiza. La IA no es un destino, sino un campo de disputa. En lugar de dejarnos arrastrar por el determinismo tecnológico o el entusiasmo ciego, es necesario construir herramientas conceptuales y políticas que nos permitan incidir en su desarrollo, orientando todos los esfuerzos hacia horizontes de justicia social, equidad cognitiva y pluralidad epistémica. En última instancia, el desafío no es puramente ético ni técnico (si algo de eso existiera), sino profundamente político: se trata de disputar el futuro mismo de la racionalidad técnica.

La convergencia de estas perspectivas subraya la complejidad de los desafíos éticos que plantea la IA. Mientras que los enfoques críticos (sean o no marxistas) advierten sobre la explotación y la desigualdad que puede generar la IA en el contexto del sistema capitalista tardío y neoliberal, el tecnooptimismo se apura en observar exclusivamente sus posibles beneficios, dejando en segundo plano el modo de producirla o si se utiliza de manera responsable. Sea como sea, es fundamental favorecer enfoques que den alerta sobre los riesgos en términos de poder y seguridad global. Estos enfoques coinciden en la necesidad de desarrollar marcos éticos y normativos que guíen el desarrollo y la implementación de la IA, en la búsqueda de que esta tecnología contribuya al bienestar humano (en las modulaciones y acepciones que esta expresión vaya tomando) y no a su detrimento.

## 2. Problemas políticos y éticos en la implementación de sistemas de IA

En esta sección se analizan los principales problemas políticos y éticos que emergen en la implementación concreta de sistemas de inteligencia artificial, proponiendo una lectura situada de las formas en que las tecnologías algorítmicas intervienen y modulan las estructuras sociales contemporáneas. En primer lugar, se examina el problema del sesgo algorítmico, entendido no solo como un defecto técnico o un error de programación, sino como una expresión de las matrices de desigualdad, discriminación y exclusión preexistentes en las sociedades (Eubanks, 2018; Noble, 2018). Lejos de ser neutrales, los sistemas de IA operan muchas veces como tecnologías de reproducción social, replicando y automatizando prejuicios vinculados a clase, género, raza o territorio. En segundo lugar, se aborda la cuestión de la responsabilidad y más tarde se trabajan las dinámicas de vigilancia, control y extracción de datos que acompañan a estos sistemas, consolidando un modelo de capitalismo informacional intensificado (Zuboff, 2019) y un régimen de gubernamentalidad algorítmica (Rouvroy & Berns, 2013) que transforma la experiencia cotidiana en información disponible para ser monetizada. Por último, la sección explora las consecuencias de estos procesos en términos de subjetivación, poniendo en diálogo aportes de la filosofía política, la teoría crítica y los estudios de ciencia y tecnología para pensar cómo la IA no solo organiza datos, sino también vidas, afectos, expectativas y posibilidades de acción. Así, esta parte del trabajo permite articular una mirada crítica sobre las tecnologías de IA, entendidas como dispositivos de poder que afectan las condiciones de existencia y amplifican –en lugar de resolver– muchas de las tensiones y desigualdades del presente.



### 2.1. El sesgo algorítmico como forma de reproducción de estructuras sociales

La IA es presentada habitualmente desde las plataformas que se benefician de ella como un artefacto técnico que promueve eficiencia, neutralidad y precisión. Sin embargo, numerosas investigaciones han demostrado que los sistemas algorítmicos no sólo no eliminan los prejuicios humanos, sino que pueden (y muchas veces tienden a) amplificarlos, automatizando formas históricas de desigualdad y replicando los

prejuicios de quienes las crean y programan. Desde la filosofía política, este fenómeno puede leerse como una cristalización tecnológica de estructuras sociales, en las que la dominación se rearticula mediante dispositivos de clasificación y predicción. Por citar un estudio relevante, Eubanks (2018) ha mostrado cómo los sistemas automatizados aplicados a la asistencia social en los Estados Unidos reproducen criterios punitivos y excluyentes hacia las poblaciones más vulnerables, especialmente a las racializadas. Algo similar ocurre con los algoritmos de vigilancia predictiva utilizados por la policía, que suelen intensificar la criminalización de ciertos grupos étnicos y territoriales (Benjamin, 2019). La IA, en estos casos, funciona como una suerte de materialización técnica del sesgo, que no hace más que institucionalizar las formas de discriminación bajo una apariencia de objetividad. En este sentido, no hay neutralidad algorítmica posible si el entrenamiento de los modelos se realiza sobre bases de datos históricamente sesgadas. La IA, lejos de eliminar el juicio humano, lo desplaza hacia etapas previas del proceso (como la selección de datos o el diseño del modelo), volviendo opaca la producción de normas y decisiones.

Desde una mirada foucaultiana, Rouvroy y Berns (2013) se han referido a procesos similares al recién aludido bajo la idea de la "gubernamentalidad algorítmica". Lo que la caracteriza es que, bajo ella, el poder no se ejerce únicamente a través de normas explícitas, sino mediante sistemas de evaluación y predicción que moldean las conductas de los sujetos sin que estos perciban la intervención del poder: "La gubernamentalidad algorítmica es pospolítica: anula el conflicto al convertirlo en problema de optimización" (Rouvroy & Berns, 2013, p. 112). Esta biopolítica automatizada produce una suerte de vida computada y automatizada, en la que una serie importante de decisiones sobre salud, educación, justicia o movilidad depende de la posición de los sujetos en matrices algorítmicas que reproducen desigualdades sociales preexistentes. Así, mientras autores como Morozov (2013, 2022) critican la "fe" en la tecnología, Rouvroy y Berns se enfocan en su funcionamiento concreto: la IA no es solo una ideología, sino una serie de mecanismos muy concretos que reconfiguran instituciones y subjetividades. Desde el punto de vista de la gubernamentalidad algorítmica, el sujeto desaparece: solo existen *clusters* de datos, volviendo al problema estructural: la IA, tal y como la conocemos, desactiva lo político y vacía de contenido la participación democrática de la ciudadanía al eliminar la deliberación y reducir todo a correlaciones.



### 2.2. Responsabilidad y agencia: ¿quién responde cuando decide una máquina?

Un segundo gran problema ético-político de la IA es el de la responsabilidad. ¿A quién se le imputa la decisión cuando esta ha sido tomada –o al menos sugerida o actuada en lo inmediato– por un sistema algorítmico? Esta cuestión adquiere especial relevancia en áreas sensibles como la medicina, la economía (los créditos, muy en particular), la justicia o la seguridad, espacios en los que los errores pueden tener consecuencias graves. Este tipo de sesgo de los sistemas de IA se manifiesta de dos maneras principales: daños de asignación y daños de representación (Eubanks, 2018; Crawford, 2021; Susskind, 2018). Los daños de asignación se refieren a la distribución injusta de recursos u

oportunidades debido a la intervención algorítmica (como, por ejemplo, en el uso de etiquetas *proxy* imprecisas, así como en la falta de transparencia en el funcionamiento y los mecanismos utilizados, que manifiestan la necesidad de considerar miradas más amplias, más allá de la simple predicción). Los daños de representación, por otro lado, se centran en cómo los algoritmos perpetúan y amplifican estereotipos existentes, afectando la percepción del mundo. Se utiliza el ejemplo de la búsqueda de imágenes de Google, donde se observa cómo las imágenes resultantes para términos como "CEO" reflejan históricamente un sesgo racial y de género. En la Internet de la automatización se encuentra, actualmente, una tensión entre representaciones precisas (que reflejan la realidad actual y, por lo tanto, sesgada) e imágenes ideales (que promueven una representación equitativa), destacando la necesidad de considerar el contexto para tomar decisiones informadas sobre qué tipo de representación es más apropiada en cada caso. Así, se vuelve evidente que es preciso comprender el contexto sociotécnico en el que operan los sistemas de IA para identificar y mitigar el sesgo algorítmico. Pero el abordaje de este problema requiere ir más allá del análisis matemático y considerar las implicaciones éticas y sociales de estas tecnologías, buscando soluciones que promuevan la justicia y la equidad.

El concepto moderno de responsabilidad se basa en la imputabilidad de una acción a un sujeto dotado de conciencia y voluntad. Pero los sistemas de IA desafían esta concepción, ya que las decisiones resultan de procesos de optimización estadística no siempre comprensibles ni por sus propios desarrolladores (Burrell, 2016). Esta situación ha dado lugar a lo que algunos autores denominan "dilución de la responsabilidad" (Coeckelbergh, 2020): una cadena de actores (programadores, empresas, usuarios, reguladores) en la que ninguno asume plenamente las consecuencias de la decisión algorítmica. Desde la filosofía política, esto plantea un problema estructural que, como ya se sugirió, podría pensarse como de despolitización: cuando nadie puede responder por una decisión, se debilita la posibilidad de exigir justicia, reparación o transformación. La IA aparece así como un nuevo dispositivo de "irresponsabilidad" que protege a los actores dominantes tras una fachada de objetividad técnica (Zuboff, 2019).



La cuestión se complejiza aún más cuando los sistemas se utilizan para "asistir" la toma de decisiones humanas, generando una zona gris de corresponsabilidad. Estudios recientes muestran que los humanos tienden a seguir las recomendaciones de la IA incluso cuando tienen dudas sobre su corrección (Green & Chen, 2019). Esto sugiere que la agencia algorítmica no es únicamente una cuestión que puede resolverse desde los protocolos, sino que transforma las relaciones sociales de autoridad, obediencia y confianza.

### 2.3. Gubernamentalidad algorítmica y nuevas formas de dominación

Más allá de los impactos técnicos o funcionales, los sistemas de IA configuran nuevas formas de poder que exigen ser pensadas en términos de dominación. La gubernamentalidad algorítmica –entendida como el uso de tecnologías automatizadas para ordenar, gestionar y anticipar comportamientos sociales– se ha convertido en un

dispositivo clave del capitalismo digital contemporáneo. En lugar de gobernar a través de leyes, se gobierna mediante infraestructuras, interfaces y modelos predictivos (Rouvroy & Berns, 2013). Uno de los aspectos más problemáticos de esta forma de poder es su opacidad estructural. Como señala Pasquale (2015), los algoritmos operan como "cajas negras" cuya lógica es inaccesible tanto para los ciudadanos como para los reguladores. Esta falta de transparencia refuerza la asimetría entre quienes controlan los sistemas y quienes son gobernados por ellos. De hecho, muchas de las decisiones que antes estaban sometidas a deliberación pública ahora se trasladan al dominio técnico, eludiendo la rendición de cuentas democrática (Crawford, 2021).

Desde una lectura gramsciana, esta lógica no solo ejerce coerción, sino que también construye consentimiento. Las plataformas digitales que implementan IA prometen conveniencia, personalización y eficiencia, moldeando los deseos y las expectativas de los usuarios en función de su racionalidad extractiva. Así, se consolida un régimen de hegemonía tecnopolítica en el que la gobernabilidad se alcanza no a través del conflicto abierto, sino mediante la integración de las subjetividades al funcionamiento técnico del sistema. Además, la expansión global de estos sistemas plantea un dilema geopolítico. Mientras algunos Estados promueven un modelo corporativo-liberal centrado en el mercado (como en los Estados Unidos), otros impulsan una versión autoritaria basada en el control estatal de la IA (como en China). En ambos casos, la IA se convierte en una tecnología de gobierno que redefine el espacio de lo público, los límites de la soberanía y el sentido mismo de lo común (Kwet, 2019).



Ante este panorama, y más allá de la regulación, cabe pensar cómo reapropiar lo común en la era algorítmica. Para ello, no basta con reclamar marcos regulatorios tradicionales, sino avanzar en el nivel político. Aunque son herramientas absolutamente necesarias, las leyes muchas veces llegan tarde y carecen de eficacia frente a la velocidad del cambio tecnológico, en especial en el caso particular de estas tecnologías que son desarrolladas con altos grados de secretismo y no publicidad (por pertenecer a empresas privadas que buscan impacto y compiten por lanzar productos antes que el resto). Además, cuando la regulación se limita a garantizar la "no discriminación" o la "privacidad de los datos", corre el riesgo de naturalizar el marco técnico dominante, sin cuestionar sus fundamentos ni su racionalidad política.

Desde una perspectiva filosófica, política y crítica, es necesario pensar la IA como un terreno de disputa que involucra modelos de subjetividad, formas de vida y regímenes de verdad muy determinados y específicos. Esto implica desplazar el debate de la mera ética técnica hacia una ética de lo común que permita imaginar y construir infraestructuras tecnológicas orientadas a fines colectivos y emancipatorios. Algunas propuestas emergentes en esta línea provienen del pensamiento del "tecnocomún" (Fuster Morell, 2020), del "tecnosocialismo" (Morozov, 2022) o de movimientos por la "soberanía algorítmica". Estas iniciativas buscan desarrollar modelos de IA de código abierto, controlados por instituciones públicas o cooperativas, con objetivos de justicia social y transparencia democrática. Pero la reapropiación política de la IA no requiere solamente buscar modificaciones en quién controla los sistemas, sino también

reimaginar qué modelos de desarrollo técnico se promueven. La discusión no puede reducirse a cómo evitar daños, sino a qué mundos queremos construir mediante estas tecnologías. Como plantea Bratton (2021), la cuestión no es simplemente regular la IA, sino repensar la forma planetaria de la política técnica, articulando infraestructuras, epistemologías y formas de vida en nuevos ensamblajes colectivos.

## 3. ¿Es posible una IA emancipadora? Límites y potencialidades

La pregunta por una inteligencia artificial emancipadora puede parecer, a primera vista, un oxímoron. ¿Cómo puede una tecnología profundamente imbricada en las lógicas del capital, el control y la predicción ofrecer posibilidades de emancipación? Sin embargo, formular esta pregunta no es un acto de ingenuidad, sino un gesto político que busca abrir el campo de lo posible más allá de su clausura tecnocrática.

### 3.1. De la crítica al diseño: repensar la IA desde otros imaginarios

La crítica a la inteligencia artificial dominante –centrada en la eficiencia, la extracción de datos y la opacidad algorítmica– ha sido esencial para desnaturalizar su aparente neutralidad (Cancela, 2024). Pero como advierte Benjamin (2019), detenerse en la crítica puede llevar a un "fatalismo tecnológico" que bloquea la imaginación política. Es necesario entonces avanzar hacia una reapropiación del diseño tecnológico, que habilite otras formas de construir y pensar la inteligencia artificial. Este enfoque implica no solo cambiar los usos de la IA, sino también los imaginarios y las futurizaciones que la sostienen. Como plantean Suchman (2011) y Costanza-Chock (2020), los sistemas técnicos son siempre performativos: encarnan valores, narrativas y supuestos sobre el mundo. Una IA emancipadora no puede nacer de los mismos marcos epistémicos que sostienen el colonialismo de datos o el capitalismo de plataformas. Requiere otros lenguajes, otras ontologías y otras epistemologías.



Aquí es donde cobran relevancia los aportes del feminismo tecnocientífico, los saberes situados y las epistemologías del sur. Estas perspectivas denuncian la universalización abstracta del "humano" en los sistemas de IA y proponen diseñar tecnologías desde la pluralidad de los cuerpos, las memorias y los territorios (Harding, 2015; Puig de la Bellacasa, 2017). En lugar de entrenar modelos para capturar patrones pasados, se trataría de imaginar infraestructuras para sostener futuros más justos y habitables. Específicamente, repasaremos de manera resumida (como para poder presentar algunas de las posibilidades recién mencionadas) cuestiones que vinculan a la IA con posibilidades de ejercer el cuidado de forma más ética y aumentar las capacidades inclusivas del mundo del trabajo.

### 3.2. Infraestructuras del cuidado y política del acompañamiento

Una de las líneas más prometedoras para pensar una IA emancipadora es la articulación entre tecnología e infraestructuras del cuidado. Frente a un modelo de

IA centrado en la optimización y el cálculo, algunos proyectos experimentan con sistemas que acompañan, asisten y sostienen, sin reemplazar la agencia humana ni suprimir la complejidad del vínculo social. Estos enfoques coinciden con una crítica al paradigma de la autonomía racional, heredado del liberalismo moderno, que subyace a muchos diseños algorítmicos. Como sostienen Puig de la Bellacasa (2017) y Mol (2008), la ética del cuidado no se basa en principios universales, sino en relaciones concretas, interdependencias situadas y afectividades materiales. Aplicada a la IA, esta perspectiva invita a imaginar sistemas que no operen desde la superioridad técnica, sino desde una lógica de copresencia, de asistencia parcial y de escucha abierta. En lugar de construir inteligencias artificiales que lo resuelvan todo, se trataría de desarrollar tecnologías que habiliten otros modos de relación: no predictivos ni extractivos, sino comprometidos con lo inacabado, lo relacional y lo frágil. Aquí, la tecnología deja de ser un sustituto del lazo social para volverse parte de su ecología, una mediación más en un entramado político de cuidado mutuo.

Esta orientación exige una reconfiguración radical de los marcos de desarrollo tecnológico que ponga en el centro no solo la eficiencia, sino también la justicia, la sostenibilidad y el deseo colectivo. En este sentido, la IA puede ser reimaginada no como una amenaza inevitable o una panacea, sino como un campo en disputa, donde es posible construir alternativas al régimen tecnoeconómico dominante.

### 3.3. Trabajo, automatización e inteligencia artificial en América Latina



La irrupción de la inteligencia artificial en el mundo del trabajo no se da de forma homogénea. En contextos latinoamericanos, marcados por estructuras económicas dependientes, informalidad laboral estructural y débiles sistemas de protección social, los impactos de la automatización adoptan características específicas que requieren ser analizadas desde un enfoque situado.

Diversos estudios (como los de CEPAL, 2022) advierten que la introducción de sistemas de IA y automatización en sectores clave como la industria manufacturera, el agro o los servicios puede intensificar las brechas sociales y productivas existentes, profundizando formas de exclusión tecnológica. En muchos casos, no se trata de una "destrucción creativa" al estilo schumpeteriano, sino de una reorganización del trabajo que precariza aún más a quienes ya estaban en situación de vulnerabilidad.

Desde una perspectiva de filosofía política, esta dinámica puede leerse como una nueva fase del colonialismo económico: no solo se extraen recursos naturales, sino también datos, tiempo y atención de poblaciones periféricas, sin participación real en los procesos de diseño, gobernanza o apropiación de la tecnología. La dependencia tecnológica se reproduce en el plano algorítmico, bajo la forma de plataformas que controlan los medios de producción digital y se imponen como infraestructuras inevitables (Katz, 2021). A esto se suma una narrativa dominante que presenta la IA como una solución mágica a los problemas del desarrollo. Bajo el lema de la "transformación digital", se promueven políticas públicas que adoptan

tecnologías de automatización sin un análisis crítico de sus efectos sociales, muchas veces bajo asesoramiento de corporaciones globales que operan con lógicas extractivistas. Así, se consolida un modelo donde la IA se convierte en un vector de acumulación por desposesión, antes que en una herramienta para la justicia social (Schwarz, 2020).

No obstante, también emergen experiencias contrahegemónicas que reimaginan el vínculo entre tecnología y trabajo desde una lógica más comunitaria. Iniciativas de cooperativismo de plataformas, tecnologías libres desarrolladas en universidades públicas, y espacios de formación tecnológica con perspectiva de género y clase, son ejemplos de una disputa activa por el sentido y el uso social de la IA en la región (Bonilla & Martínez, 2021). Pensar una IA emancipadora en América Latina exige, por tanto, articular lucha política, soberanía tecnológica y justicia laboral. Se trata de rechazar tanto el determinismo tecnofóbico como el tecnooptimismo acrítico, y apostar por una construcción situada de saberes y herramientas que respondan a las necesidades, memorias y horizontes colectivos de nuestros pueblos.

## Conclusiones: pensar con y contra la inteligencia artificial

Enfrentar críticamente el desarrollo contemporáneo de la inteligencia artificial no implica rechazarla en bloque ni abrazarla de forma acrítica. Tanto la tecnofobia como el solucionismo tecnoentusiasta constituyen formas de clausura del pensamiento. Una mirada filosófica-política comprometida con el presente debe más bien sostener la tensión, el intersticio, la ambivalencia: pensar con y contra la IA. La opacidad de muchos modelos avanzados de IA, particularmente en *deep learning*, dificulta la auditoría externa y el derecho a explicación. Cuando un algoritmo niega un crédito o recomienda una sentencia judicial, los afectados merecen comprender los motivos detrás de esa decisión. Esto ha llevado al desarrollo de enfoques como algoritmos interpretables y a regulaciones que exigen mayor transparencia en sistemas de alto riesgo. La trazabilidad se convierte así en un requisito ético fundamental.



El impacto ambiental de la IA representa otra dimensión crítica que el debate ético deberá tomar en consideración lo antes posible. El entrenamiento de modelos complejos consume cantidades masivas de energía, con huellas de carbono comparables al consumo anual de decenas de hogares. Paralelamente, la concentración del desarrollo tecnológico en pocas corporaciones y pocos países genera desigualdades globales, marginando a regiones en desarrollo. Abordar estos desafíos requiere inversión en IA sostenible y cooperación internacional para democratizar el acceso a estos avances.

Principios como justicia, transparencia, responsabilidad, privacidad y sostenibilidad deben guiar el desarrollo de la IA. Las iniciativas actuales buscan establecer marcos éticos que eviten discriminación, garanticen explicabilidad, definan líneas claras de responsabilidad, protejan datos personales y reduzcan el impacto ambiental. La participación diversa de filósofos, ingenieros, legisladores y sociedad civil resulta esencial

para crear sistemas alineados con el bien común. El ritmo acelerado de la innovación exige mecanismos regulatorios ágiles que equilibren progreso tecnológico con protección de derechos fundamentales, aprendiendo de errores pasados para construir un futuro donde la IA sirva a toda la humanidad de manera equitativa y sostenible.

Lejos de ser una herramienta neutra, la IA condensa las coordenadas del orden social contemporáneo: captura, predicción, control, aceleración, eficiencia. Su funcionamiento actual no es ajeno al modo de producción capitalista ni a las estructuras de dominación que atraviesan las subjetividades, los territorios y los imaginarios. Por ello, cualquier ética de la IA que no aborde estas dimensiones estructurales corre el riesgo de volverse decorativa, un suplemento de legitimación en lugar de una intervención transformadora.

A la vez, renunciar a la pregunta por una IA emancipadora sería ceder el campo tecnológico al diseño hegemónico. En lugar de cerrar la cuestión, es preciso abrirla. ¿Qué otras inteligencias son posibles? ¿Qué otras formas de relación entre técnica, cuerpo y política podemos imaginar? En este sentido, resulta particularmente fértil retomar las propuestas de Hui (2021) en torno a la "tecnodiversidad", entendida como la posibilidad de pensar múltiples cosmologías técnicas no reducidas al universalismo tecnocientífico occidental. Sin embargo, para que esta noción no se vuelva un gesto meramente antropológico o culturalista, es necesario repolitizarla desde coordenadas regionales. Pensar la tecnodiversidad desde América Latina implica interrogar nuestras propias genealogías técnicas, nuestras formas de relación con la naturaleza, la comunidad y el tiempo, así como también con los legados coloniales que configuran nuestras infraestructuras y subjetividades.



La ética de la IA no debe reducirse a un código de buenas prácticas ni a un marco regulatorio mínimo. Debe pensarse como un campo de intervención en el que se juega el sentido mismo de lo humano, lo vivible y lo común en el siglo XXI. Así, una ética latinoamericana de la IA debería superar la oposición binaria entre adopción subordinada o rechazo nostálgico, e inscribirse en una cartografía de luchas por la soberanía tecnológica, la justicia cognitiva y la reapropiación del diseño. Se trata de asumir que la IA no es solo un objeto de análisis, sino también un terreno de disputa: una mediación que puede contribuir a reproducir la lógica del capital o a habilitar formas inéditas de politicidad, cuidado y cohabitación.

La IA es una tecnología que permite a las máquinas realizar tareas que normalmente requieren (o requerían) la intervención directa e inmediata de la inteligencia humana, ya que, en lugar de seguir instrucciones específicas, pueden analizar datos, reconocer patrones y tomar decisiones basadas en información que se les provee o captan. Esto les permite realizar una amplia gama de acciones de manera autónoma y mejorar su rendimiento con el tiempo. Hoy estamos rodeados de una infinidad de inteligencias artificiales "estrechas", diseñadas para tareas específicas como el reconocimiento de voz, la traducción de textos o discursos, la detección de fraudes, la puesta en acto de muchas terapias preventivas y la recomendación de productos, entre muchas otras. La IA se basa en algoritmos y modelos matemáticos que permiten a las máquinas aprender de datos históricos y tomar decisiones

basadas en patrones identificados. Esto puede ser beneficioso en muchos aspectos de la sociedad como la atención médica, la movilidad autónoma, la investigación científica y la automatización de procesos industriales, ya que puede mejorar la eficiencia, la precisión y la toma de decisiones. Sin embargo, la IA también plantea desafíos importantes como la privacidad, la seguridad, la discriminación y la responsabilidad. Es fundamental establecer regulaciones y directrices para garantizar que la IA se utilice de manera responsable y en beneficio de toda la sociedad. Pero también es esencial el acompañamiento ético y el posicionamiento político de las sociedades frente a estos avances, para no permitir que sus beneficios solo favorezcan a sectores muy reducidos.

Nos encontramos, entonces, frente a dos grandes desafíos. Por un lado, debemos investigar cómo controlar el desarrollo de tecnologías que pueden ser útiles para la sociedad sin que eso implique frenarlas. ¿A qué sectores impulsar y a cuáles observar más de cerca? ¿Cómo hacer que los beneficios se repartan más equitativamente en toda la sociedad y no se conviertan en más concentración de poder para élites muy pequeñas? Para eso es fundamental el ejercicio de la prudencia ética y la socialización de la educación digital.

Por el otro lado –y sobre todo–, recordar que el algoritmo es también una relación social: implica una mutación general de la subjetividad. Las interacciones digitales son mediaciones que se ocultan en su carácter medial, de modo que recuperar la posibilidad de institucionalizar de algún modo las relaciones mediadas por contextos digitales, y especialmente aquellos gestionados con o a través de IA, es fundamental. Por eso, como dicen Pasquinelli y Joler (2021), hay que tener presente que una de las mayores opacidades de la IA es ocultar su carácter político.



**Reconocimiento de uso de IA**

El autor de este artículo declara que ha utilizado herramientas de inteligencia artificial para tareas auxiliares como corregir, resumir o editar ciertas frases o párrafos determinados, así como buscar ciertos sinónimos específicos con el fin de ganar claridad o fluidez sin alterar el contenido. También se ha utilizado IA para revisar que el estilo de las referencias bibliográficas estuviera correctamente aplicado. No se ha utilizado estas herramientas para tomar decisiones metodológicas ni para generar los argumentos o su hilo conductor.

**Bibliografía**

Álvarez Álvarez, A., Arroyo Prieto, L., Avila Pinto, R., Crusellas Tura, E., Vázquez, M., Martínez Portell, L., Mata i Solsona, M., Ruiz de Alda, T., Ruiz Costa-Jussà, M. & Torras Genís, C. (2022). Datos abiertos y la inteligencia artificial. Barcelona: Generalitat de Catalunya.

Ávila, R., Brun, L., Bull, B., Cecchini, S., Costafreda, A., Rivera, L., Peirano, M., Sanahuja, J. & Svampa, M. (2023). La triple transición. Visiones cruzadas desde Latinoamérica y la Unión Europea. Madrid: Fundación Carolina & Oxfam Intermón.

Belli, L. & Zingales, N. (2022). Data protection and artificial intelligence inequalities and regulations in Latin America. Computer Law & Security Review, 47, 105761. DOI: https://doi.org/10.1016/j.clsr.2022.105761.

Benkler, Y. (2006), The Wealth of Networks: how social production transforms markets and freedom. New Haven: Yale University Press.

Benjamin, R. (2019). Race After Technology: Abolitionist Tools for the New Jim Code. Cambridge: Polity Press.

Bratton, B. H. (2021). The Revenge of the Real: Politics for a Post-Pandemic World. Londres & Nueva York: Verso Books.

Brynjolfsson, E. & McAfee, A. (2014). The Second Machine Age. Nueva York: W. W. Norton.

Bonilla, Y. & Martínez, G. (2021). Tecnologías comunitarias y soberanía digital: experiencias desde América Latina. Comunicación y Sociedad, 38(1), 1-24.



Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 1–12. DOI: https://doi.org/10.1177/2053951715622512.

Cancela, E. (2024). Utopías digitales. Imaginar el fin del capitalismo. Buenos Aires: Prometeo.

CEPAL (2022). El futuro del trabajo en América Latina y el Caribe en el contexto de la digitalización y la automatización. Comisión Económica para América Latina y el Caribe. Recuperado de: https://www.cepal.org.

Coeckelbergh, M. (2020). AI Ethics. Cambridge: MIT Press.

Cortina, A. (2024). ¿Ética o ideología de la inteligencia artificial? Madrid: Tecnos.

Costanza-Chock, S. (2020). Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge: MIT Press.

Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press.

Danesi, C. (2025). Lo que dicen los primeros estudios sobre las relaciones de los humanos con los chats de IA: pueden aliviar la soledad, pero también aislar y generar dependencia.

El País, 1 de abril. Recuperado de: https://elpais.com/tecnologia/2025-04-01/lo-que-dicen-los-primeros-estudios-sobre-las-relaciones-de-los-humanos-con-los-chats-de-ia-pueden-aliviar-la-soledad-pero-tambien-aislar-y-generar-dependencia.html.

de Sousa Santos, B. (2017). Justicia entre saberes: Epistemologías del Sur contra el epistemicidio. Madrid: Morata.

D'Ignazio, C. & Klein, L. (2020). Data Feminism. Cambridge: MIT Press.

Dyer-Witheford, N. (2015). Cyber-Proletariat: Global Labour in the Digital Vortex. Londres: Pluto Press.

Dyer-Witheford, N., Kjøsen, A. N. & Steinhoff, J. (2024). Poder inhumano. Inteligencia artificial y el futuro del capitalismo. Buenos Aires: Prometeo.

Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. Nueva York: St. Martin's Press.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, Ch., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. & Vayena, E. (2018). AI4People. An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28, 689-707.



Fraser, N. (2008). Scales of Justice. Nueva York: Columbia University Press.

Fuster Morell, M. (2020). The Commons as a Digital Public Infrastructure: Governance Design for an Alternative Internet. P2P Models Project. Recuperado de: https://p2pmodels.eu.

García Serrano, A. (2016). Inteligencia Artificial, fundamentos práctica y aplicaciones. RC.

Gray, M. L. & Suri, S. (2019). Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass. Boston: Houghton Mifflin Harcourt.

Green, B. & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (pp. 90-99). DOI: https://doi.org/10.1145/3287560.3287573.

Harding, S. (2015). Objectivity and Diversity: Another Logic of Scientific Research. Chicago: University of Chicago Press.

Hui, Y. (2021). Art and Cosmotechnics. Mineápolis: University of Minnesota Press.

Katz, R. (2021). Transformación digital y brechas estructurales en América Latina. Fundación Telefónica.

Krauss, P. (2023). Künstliche Intelligenz und Hirnforschung. Neuronale Netze, Deep Learning und die Zukunft der Kognition. Springer.

Kurzweil, R. (2005). The Singularity is Near: When Humans Transcend Biology, Nueva York: Penguin.

Kwet, M. (2019). Digital colonialism: US empire and the new imperialism in the Global South. Race y Class, 60(4), 3-26. DOI: https://doi.org/10.1177/0306396818823172.

Lupton, D. (2016). The quantified self: A sociology of self-tracking. Cambridge: Polity Press.

Mejías, U. A. & Couldry, N. (2019). Colonialismo de datos: repensando la relación de los datos masivos con el sujeto contemporáneo. Virtualis, 10(18), 78-97. Recuperado de: https://www.revistavirtualis.mx/index.php/virtualis/article/view/289.

Mohamed, S., Png, M. T. & Isaac, W. (2020). Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. Philosophy & Technology, 33, 659-684. DOI: https://doi.org/10.1007/s13347-020-00405-8.

Mol, A. (2008). The Logic of Care: Health and the Problem of Patient Choice. Routledge.



Moor, J. H. (1985). What is Computer Ethics? Metaphilosophy, 16(4), 266-275.

Morozov, E. (2013). To Save Everything, Click Here. Nueva York: PublicAffairs.

Morozov, E. (2022). It's the Tech, Stupid: On the Political Economy of Digital Capitalism. Londres & Nueva York: Verso Books.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. Nueva York: New York University Press. DOI: https://doi.org/10.18574/nyu/9781479833641.001.0001.

Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Cambridge: Harvard University Press.

Pasquinelli, M. & Joler, V. (2021). El Nooscopio de manifiesto. La Fuga, 25. Recuperado de: http://2016.lafuga.cl/el-nooscopio-de-manifiesto/1053.

Puig de la Bellacasa, M. (2017). Matters of Care: Speculative Ethics in More than Human Worlds. Mineápolis: University of Minnesota Press.

Ricaurte, P. (2019). Data Epistemologies, The Coloniality of Power, and Resistance. Television & New Media, 20(4), 350-365. DOI: https://doi.org/10.1177/1527476419831640.

Rouvroy, A. & Berns, T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation. Le disparate comme condition d'individuation par la relation? Réseaux, 177, 1/2013, 163-196.

Schneider, S. (2019). Artificial You: AI and the Future of Your Mind. Princeton: Princeton University Press.

Serrano, M. & Peletier, I. (2024). En qué piensan los robots. Guía definitiva para entender la disrupción que viene. Madrid: Editorial Tecnología y Sociedad.

Suchman, L. (2011). Anthropological Relocations and the Limits of Design. Annual Review of Anthropology, 40, 1-18.

Torres, J. (2023). La inteligencia artificial explicada a los humanos. Barcelona: Plataforma Actual.

Spitzer, M. (2023). Künstliche Intelligenz. Dem Menschen überlegen wie künstliche Intlligenz uns rettet und bedroht. Droemer.

Schwarz, A. (2020). Data colonialism and algorithmic capitalism: New territories of extraction and control. Media, Culture & Society, 42(6), 855-872. DOI: https://doi.org/10.1177/0163443719876541.



Susskind, J. (2018). Future Politics: Living Together in a World Transformed by Tech. Oxford: Oxford University Press.

UNESCO (2021). Recomendación sobre la ética de la inteligencia artificial. Recuperado de: https://bit.ly/4j85WTr.

Veraza Urtuzuástegui, J. (2024). Karl Marx, la inteligencia artificial y el gobierno despótico de la producción. México: Editorial Crítica.

Xun, J. (2025). "Hipnocracia" o el régimen de la sociedad adormecida con dos sumos "sacerdotes": Trump y Musk. El País, 26 de marzo. Recuperado de: https://elpais.com/tecnologia/2025-03-26/hipnocracia-o-el-regimen-de-la-sociedad-adormecida-con-dos-sumos-sacerdotes-trump-y-musk.html.

Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Nueva York: PublicAffairs.

# Experta ignorancia. Analfabetismo humanista y desconexión disciplinar en la ética de la inteligencia artificial

# Ignorância de especialistas. Analfabetismo humanista e desconexão disciplinar na ética da inteligência artificial

# *Expert Ignorance. Humanist Illiteracy and Disciplinary Disconnection in the Ethics of Artificial Intelligence*

**Camilo Matías Quilapán y Verónica Moreno** *



En respuesta a las profundas consecuencias de la implementación de la inteligencia artificial (IA) en la vida humana, numerosos académicos han tornado su reflexión experta a lo que se ha llamado la ética de la inteligencia artificial, una disciplina que agrupa múltiples sectores del saber en torno al objetivo común de reducir los riesgos asociados a esta tecnología e, idealmente, potenciar su aporte positivo. Sin embargo, a pesar del esfuerzo colectivo que se está llevando a cabo, se puede ver aún una suerte de disociación entre expertos técnicos y humanistas. El presente artículo evalúa el rol social de la academia humanista y el fenómeno del razonamiento práctico periférico (RPP), para luego proponer la alfabetización recíproca bicondicional (ARB) en las técnicas y desafíos de la IA. Finalmente, se presenta una propuesta con los contenidos mínimos para promover la alfabetización en IA entre los académicos humanistas para permitir una verdadera ética práctica sobre la IA.

**Palabras clave:** ética IA; alfabetización en IA; diálogo interdisciplinario; desarrollo responsable; razonamiento práctico periférico

* *Camilo Matías Quilapán:* académico y docente de filosofía, Escuela de Humanidades, Departamento de Formación Integral, Universidad San Sebastián, Chile. Miembro del Grupo de Estudio e Investigación Ética, IA y Desafíos Actuales, Chile. Correo electrónico: camilo.matias@uss.cl. ORCID: https://orcid.org/0009-0004-1574-3795. *Verónica Moreno:* investigadora de desafíos éticos y epistemológicos de la IA. Candidata a doctora en filosofía por la Pontificia Universidad Católica de Chile. Correo electrónico: ve.moreno@estudiante.uc.cl. ORCID: https://orcid.org/0009-0004-4999-9908.

Em resposta às profundas consequências da implementação da inteligência artificial (IA) na vida humana, vários acadêmicos voltaram sua reflexão especializada para o que foi chamado de ética da inteligência artificial, uma disciplina que reúne vários setores do conhecimento em torno do objetivo comum de reduzir os riscos associados a essa tecnologia e, idealmente, potencializar suas contribuições positivas. Entretanto, apesar do esforço coletivo realizado, ainda é possível observar uma espécie de dissociação entre especialistas técnicos e humanistas. Este artigo avalia o papel social da bolsa de estudos humanística e o fenômeno do raciocínio prático periférico (RPP) e, em seguida, propõe a alfabetização recíproca bicondicional (ARB) nas técnicas e desafios da IA. Por fim, é apresentada uma proposta com conteúdo mínimo para promover a alfabetização em IA entre acadêmicos humanísticos para permitir uma abordagem ética verdadeiramente prática à IA.

**Palavras-chave:** ética da IA; alfabetização em IA; diálogo interdisciplinar; desenvolvimento responsável; raciocínio prático periférico

*In response to the profound consequences of the implementation of artificial intelligence (AI) in human life, numerous scholars have turned their expert reflection to what has been called AI ethics. This discipline brings together multiple sectors of knowledge around the common goal of reducing the risks associated with this technology and, ideally, enhancing its positive contributions. However, despite the collectiveness of the effort, a sort of dissociation between technical experts and humanists remains. This article evaluates the social role of humanist academia and the phenomenon of peripheral practical reasoning (PPR), proposing in turn the biconditional reciprocal literacy (BRL) regarding the techniques and challenges of AI. Lastly, a proposal with the minimum contents required for promoting AI literacy within humanist scholars, to enable a truly ethical practice regarding AI, is presented.*

***Keywords:*** *AI ethics; AI literacy; interdisciplinary dialogue; responsible development; peripheral practical reasoning*

## Introducción

Los últimos tres años hemos sido partícipes de una nueva revolución tecnológica representada en las mentes de muchos por los avances de ChatGPT. Desde 2022 se ha visto cómo esta herramienta ofrece un nivel de conversación inaudito y continúa mejorando en ámbitos como la creación de documentos, escritura de código, generación de imágenes y asistencia en tareas. Las consecuencias de la implementación de esta tecnología son profundas y abarcan todos los ámbitos de la vida humana. En respuesta, numerosos académicos han tornado su reflexión experta a lo que se ha llamado la ética de la inteligencia artificial, una disciplina que agrupa múltiples sectores del saber –científicos de datos, economistas, sociólogos, abogados, filósofos y lingüistas, entre otros– en torno al objetivo común de reducir los riesgos asociados a esta tecnología e, idealmente, potenciar su aporte positivo. Así, se ha visto un incremento en los artículos sobre regulación de IA, principios éticos para su desarrollo e implementación, vías para el desarrollo sostenible y responsabilidad social y corporativa, y advertencias sobre las consecuencias sociales de algoritmos sesgados y abusos de estas herramientas.

Ahora bien, el concepto de "inteligencia artificial" (IA) es amplio e incorpora tecnologías muy diferentes, tanto respecto de su estructura como propósitos. Para explicar esto, la introducción del libro *AI Snake Oil* (Narayanan & Kapoor, 2024) usa una analogía fantástica que merece consideración. El término "inteligencia artificial" agrupa tantas herramientas distintas como el término "vehículo". Si se consideran los vehículos sin distinguir sus tipos, es difícil evaluar los problemas técnicos, éticos, sociales y legales que implican. Se pueden publicar estudios sobre emisiones de dióxido de carbono, uso de recursos fósiles, peligro de asfixia, riesgo de colisión y muchas otras amenazas. Sin embargo, estas amenazas no se dan siempre en cada tipo de vehículo ni de la misma manera. Un auto y un avión no emiten la misma cantidad de dióxido de carbono y son diferentes los riesgos de asfixia por fallas en un submarino y en un cohete espacial. Más aún, hay vehículos como la bicicleta que no contaminan y ayudan a promover la salud de las personas, aunque son riesgosas en caso de colisión. De la misma manera, a la hora de reflexionar sobre la IA es crucial tener un mínimo conocimiento de su funcionamiento, tipos y contextos de aplicación, pues los beneficios y riesgos cambian drásticamente. Para analizar adecuadamente las implicaciones sociales, éticas y legales de la IA, se necesita un conocimiento técnico mínimo que vaya más allá del uso cotidiano. Aunque forma parte del día a día, es imposible aportar críticamente a la discusión sin una comprensión profunda y bien fundamentada de esta tecnología. Esta es la llamada alfabetización en IA, un subtipo de alfabetización digital que incorpora tanto conocimiento teórico como manejo práctico de la inteligencia artificial.



A pesar del esfuerzo colectivo que se está llevando a cabo para desarrollar reflexiones sobre la eticidad de distintos fenómenos ligados a la IA, se puede ver aún una suerte de disociación en la academia, donde por un lado avanzan los expertos técnicos y por otro los humanistas. Mientras los primeros se enfocan en desarrollar modelos computacionalmente más eficientes, transparentes y con data equilibrada, los

segundos se centran en los cambios que está sufriendo la sociedad en, por ejemplo, los sistemas educativos y laborales. Más aún, no solo hay una desconexión entre investigadores técnicos y humanistas, sino entre ambos y la sociedad en general, ya que los esfuerzos por dar lugar a una IA segura están lejos de dialogar con los temores y las necesidades de las personas afectadas por estas herramientas.

Esta desconexión entre el enfoque técnico y humanista lleva consigo una escasez de diálogo y de lenguaje común. Esto se puede ver, por ejemplo, en la manera completamente disímil de entender imparcialidad (*fairness*) y sesgo (*bias*) en el mundo de la programación y en el de las humanidades. En el primero se refiere al equilibrio y desequilibrio de un algoritmo o base de datos, mientras que en el segundo estas palabras aluden a equidad y prejuicios sociales. A su vez, la escasez de lenguaje común entorpece el diálogo para el desarrollo responsable de la IA.

Más aún, así como las técnicas de *machine learning* se perfeccionan y se usan para resolver preguntas humanas, así también los dilemas sociales se convierten cada vez más en dilemas técnicos. De esta manera, dado que el estudio de la ética aplicada requiere tanto consideraciones teóricas como conocimientos de las condiciones concretas de aplicación, es crucial que la reflexión ética vaya de la mano con un adecuado dominio de las disciplinas técnicas. Frente a ello, destacamos la necesidad de promover la alfabetización en IA entre los académicos humanistas.



En este artículo se aborda en esta disociación de la experiencia y la importancia de la alfabetización en IA. Para ello, la segunda sección describe el rol de la academia humanista en la sociedad, seguida por la presentación del fenómeno de disociación en términos de la razón práctica periférica (RPP) en la tercera sección. En respuesta, la cuarta sección propone un camino fundamental para generar colaboración interdisciplinaria: la alfabetización recíproca bicondicional (ARB). Ésta implica a la vez una asistencia humanista al desarrollo técnico para dar lugar a un alineamiento valórico y un esfuerzo de alfabetización en IA entre los académicos de humanidades. La sección final propone una serie de distinciones a la hora de hablar de IA que conforman los requerimientos mínimos de conocimiento para generar un análisis atingente y significativo sobre los fenómenos sociales que responden al avance de la IA.

## 1. La tarea de la academia humanista frente a la IA

La relación entre humanidades y tecnología, específicamente con la inteligencia artificial, tiene un antecedente histórico en la distinción entre humanidades y ciencia. En el siglo XVII, Newton y otros aún se identificaban como “filósofos naturales”, reflejo de un tiempo en que ciencia y filosofía no estaban separadas. No obstante, en el siglo XIX esta división se torna evidente: en *The Philosophy of the Inductive Sciences*, Whewell (1967 [1840]) introduce el término “científico” en un sentido que excluye a los filósofos, marcando el surgimiento de una noción de “ciencia” diferenciada –y a veces contraria– a las humanidades.

Snow propuso la noción de “dos culturas” (1961) para evidenciar la creciente separación entre ciencias y humanidades. Según él, existen dos grupos polarizados: los intelectuales literarios, que se autodenominan intelectuales excluyendo a otros saberes, y los científicos, representados especialmente por los físicos. Esta división ha generado mutua incomprensión y en ocasiones abierta animadversión, resultando en círculos herméticos cuyos trabajos, marcados por lenguajes técnicos propios, se vuelven inaccesibles incluso entre ambos polos.

Ya en la segunda edición de su opúsculo (1963), Snow propuso la idea de una “tercera cultura” como espacio de convergencia entre ciencias y humanidades, dos polos cada vez más distanciados. Sin embargo, fueron científicos como Einstein, Sagan o Hawking quienes, con su labor divulgativa, comenzaron a materializar esta tercera cultura, acercando la ciencia al público general y saliendo del círculo hermético, aunque no sin cierta desconfianza de parte de colegas más puros. En *The Third Culture* (1995), Brockman argumenta que el ámbito intelectual estadounidense estuvo dominado durante años por sectores literarios reaccionarios, contrarios al avance científico, alejados del empirismo y encerrados en jergas autorreferenciales. A su juicio, estos “intelectuales literarios” abandonaron el espacio público, cediendo el lugar de la tercera cultura a científicos con vocación de impacto social, quienes, motivados por la relevancia de su trabajo, asumieron el rol de “intelectuales contemporáneos”, destacándose en la divulgación incluso más que en la investigación estricta.



En el contexto actual marcado por el fenómeno de la IA y sus múltiples implicancias, reaparece la necesidad de una tercera cultura, en la que confluyan tanto los desarrolladores técnicos de IA como quienes analizan sus impactos sociales, éticos y jurídicos. Esta posibilidad se basa en la persistencia de dos polos opuestos: por un lado, los especialistas tecnocientíficos encargados del desarrollo técnico; por el otro, los humanistas que abordan críticamente sus efectos. Al igual que en la división descrita por Snow, ambos polos conviven dentro de la misma academia, lo que impone a las humanidades un desafío doble: reflexionar e investigar con una auténtica mirada crítica los fenómenos vinculados a la IA, y al mismo tiempo dialogar con los sectores que configuran su contexto social. Dicha mirada crítica, entendida en su sentido originario griego (κρίνω), implica separar, distinguir, juzgar, interpretar y resolver los fenómenos que se estudian. Aplicado a la IA, esto exige abordar con rigor sus implicancias sociales y éticas –por ejemplo, en educación–, considerando que actores como OpenAI tienen motivaciones diferentes a las de los educadores (Popenici, 2023). Para ello, la academia humanista debe conocer los componentes técnicos básicos del fenómeno que estudia, como la distinción entre IA generativa e IA predictiva. Esta comprensión es una exigencia inherente al pensamiento crítico y condición para el segundo gran desafío: hacer dialogar no solo a los polos técnicos y humanistas, sino también a los grandes sectores de la sociedad.

Para efectos prácticos, dividimos en tres grandes áreas aquellos grupos con los que la academia humanista tiene la exigencia de dialogar: academia e industria tecnocientífica, ciudadanía y política. En el primer grupo se ubican tanto académicos de las humanidades

como de las ciencias, junto con expertos técnicos del ámbito industrial que muchas veces provienen de la misma academia. El caso de OpenAI es ilustrativo: aunque independiente institucionalmente, muchos de sus miembros se formaron en universidades de prestigio como Toronto, Berkeley, NYU o MIT. Iniciativas como NextGenAI (Jauregui-Volpe, 2025), que integran a más de 15 centros de investigación, muestran que el liderazgo en esta tercera cultura sigue en manos de científicos, relegando nuevamente a las humanidades, como ya anticipaba Brockman. En cuanto a la ciudadanía, se entiende aquí a quienes, sin poseer experticia técnica, son los principales receptores de los efectos del desarrollo de la IA. Como muestra, los trabajos de Hinton han permitido mejoras en la detección temprana del cáncer de mama (McKinney *et al.*, 2020), con impacto directo en los ciudadanos, evidenciándose estos precisamente como el grupo receptor. Por último, la política se comprende –siguiendo a Yáñez (2019)– como una actividad moral orientada al bien común. En este sentido, quienes ejercen cargos públicos no solo deben emplear herramientas como la IA, sino también comprender sus efectos, por ejemplo, en iniciativas de salud pública como la aplicación de la IA al diagnóstico y tratamiento de la epilepsia (Lucas *et al.*, 2024). Dado su impacto potencial tanto positivo como negativo, la IA es un asunto que involucra directamente a quienes tienen el deber de gobernar con prudencia y orientar sus decisiones hacia el bien común.

La exigencia de diálogo entre los tres grandes sectores –academia e industria, política y ciudadanía– deriva del hecho de que esta última representa el terreno común donde confluyen los efectos del desarrollo y aplicación de la IA. Los ciudadanos son, en efecto, los principales receptores de las consecuencias del uso de estas tecnologías, aunque no participen directamente en su construcción o implementación, no por desinterés, sino por la falta de experticia técnica, cuyo acceso exige una inversión significativa de tiempo y recursos. Así, el ciudadano común –ajeno tanto al ámbito académico como al político– depende de la información y decisiones de expertos y autoridades, ya sea en elecciones, consultas públicas u otros espacios de deliberación. Por lo tanto, es el propio bien común ciudadano el que impone la necesidad de diálogo entre los sectores implicados. Y dado que este diálogo no se limita a lo técnico, sino que involucra saberes de la ética aplicada, la sociología o la epistemología, se requiere que el académico humanista, antes de formular cualquier propuesta, realice una reflexión crítica rigurosa, algo que no puede llevarse a cabo sin una alfabetización mínima en IA.



En síntesis, la relación entre humanidades, ciencia y tecnología ha estado históricamente atravesada por una separación creciente, como evidenció Snow al proponer la noción de las dos culturas. Esta fractura persiste en el contexto actual de la IA, donde técnicos y humanistas operan desde lógicas distintas, incluso dentro de la misma academia. En respuesta, Brockman propuso la idea de una tercera cultura, en la que científicos con vocación divulgativa asumieron el rol intelectual que los humanistas dejaron vacante, articulando un puente entre el saber técnico y la sociedad. En este escenario, las humanidades enfrentan un doble desafío: por un lado, desarrollar una mirada crítica rigurosa que permita analizar las implicancias éticas y sociales de la IA; por otro, convertirse en un puente de diálogo entre los tres sectores clave: academia tecnocientífica-industria, ciudadanía y política. Mientras la industria lidera el desarrollo

desde instituciones técnicas, la ciudadanía recibe los impactos –como en los casos de aplicación de IA al cáncer de mama y la epilepsia–, y la política debe orientar su uso al bien común. La reactivación del rol de las humanidades en esta tercera cultura requiere una alfabetización técnica mínima, exigida tanto por la coherencia de su propia mirada crítica como por el bien común ciudadano, que constituye el punto de convergencia entre los sectores académico y político. No obstante, esta alfabetización y esta orientación al bien común se ven hoy obstaculizadas por un problema estructural de base que condiciona el diálogo entre los sectores y la eficacia crítica de sus intervenciones.

## 2. Un problema estructural de base: el razonamiento práctico periférico (RPP)

Las tres áreas que deberían entrar en diálogo tienen un problema estructural que condiciona e, incluso, imposibilita un auténtico diálogo entre las partes involucradas respecto a la construcción, implementación, uso y efectos de la IA. Esto, a su vez, impide la alfabetización y afecta, finalmente, al bien común de la ciudadanía. A este problema estructural lo denominamos, como ya se mencionó, razonamiento práctico periférico (RPP).

El razonamiento, en su sentido más clásico, consiste en discurrir, mediante juicios, hacia una determinada conclusión; conclusión que es fruto de los contenidos y condicionantes dados en las premisas previas. Así, por ejemplo, de la premisa "Todos los hombres son mortales" puedo pasar a "Sócrates es hombre" y, gracias a que hay un término que sirve de puente entre ambas, a saber, "hombre", puedo concluir que Sócrates es mortal. Ahora bien, cuando hablamos de razonamiento práctico, siguiendo la línea aristotélica, nos referimos a aquel cuyo fin no es el saber, sino más bien el obrar. Este obrar requiere conocimiento y reflexión previa, pero su finalidad última es la acción concreta. De este modo, entramos en el campo de la ética, puesto que la conclusión de un razonamiento práctico no es una proposición teórica, sino una acción voluntaria, que puede evaluarse según su ajuste o desajuste respecto a las premisas previamente conocidas. En este sentido, aunque el fin sea crucial, también lo es la deliberación previa, ya que ésta, al establecer lo que es necesario y verdadero predicar en una situación concreta, permite que el agente prudente actúe en consecuencia con el juicio racionalmente elaborado (Trujillo, 2007). Así, a grandes rasgos, un razonamiento práctico adopta una estructura en tres partes: primero, se formula una regla de conducta general o universal (premisa mayor); segundo, se reconoce que un agente se halla ante una situación concreta comprendida bajo esa norma (premisa menor); y, como conclusión, el agente actúa de forma prudente y coherente con ambas premisas, ejecutando la acción que corresponde según la norma reconocida.

Ahora bien, un razonamiento práctico puede ser erróneo por una causa particular, aunque no la única, a saber, por ignorancia. La ignorancia puede causar un error de conciencia. Para articular un sentido del error de conciencia que atienda a nuestro análisis, debemos tener presente lo siguiente: el acto y el silogismo práctico previo que opera como deliberación pueden ser analizados en primera persona o,

por contraparte, desde la perspectiva de tercera persona. Salvo el incontinente u otros casos de vicio moral, desde una perspectiva de primera persona, el sujeto que delibera y obra supone, en su deliberación –o sea, en las premisas que constituyen su razonamiento práctico–, el cumplimiento mínimo de patrones de racionalidad interna que sustenten su creencia. Esto implica que debe haber una consistencia interna entre los deseos, fines y acciones del agente que obra. El caso del incontinente, que dejamos de lado en este trabajo, podría ubicarse entre aquellos que presentan irracionalidad interna, por cuanto, por ejemplo, su acción no se adecúa con los deseos o fines del mismo agente (Vigo, 2013). En cambio, en el error de conciencia, el sujeto que obra bajo este error supone la racionalidad interna de sus postulados y su acción. Sin embargo, puede luego darse cuenta de su error por un factor esencial respecto a la constitución del error mismo; a saber, que darse cuenta del error supone haber salido de su ilusión.

Si en el ámbito interno es la subjetividad la que se encuentra envuelta en el *velo del error* al punto de considerar que sí se están cumpliendo los patrones de racionalidad interna, son sin embargo los patrones de racionalidad externa los que dan cuenta del vacío por ignorancia en el aspecto interno. Aquí pasamos a la perspectiva de tercera persona, donde la racionalidad externa apunta a determinar, enjuiciar y valorar de modo más riguroso el contenido proposicional de los deseos, medios y fines involucrados en el razonamiento práctico. Más en concreto, la racionalidad externa permite sopesar y determinar con criterios de valor de carácter intersubjetivo (Vigo, 2013) los contenidos de la premisa mayor y la premisa menor del razonamiento práctico. Esta distinción es clave para evaluar críticamente no solo las acciones, sino la estructura misma del juicio práctico en situaciones complejas como las involucradas en decisiones técnicas respecto a la implementación y el uso de IA con impacto social.



Pues bien, entonces ¿qué es el RPP? Es un tipo de razonamiento práctico en el cual el agente, aunque cumple con los requisitos de racionalidad interna (coherencia entre fines, medios y acción), falla en términos de racionalidad externa, debido a que carece del conocimiento o la experticia necesaria sobre el campo específico en el que pretende actuar. Por ello, su razonamiento permanece cognoscitivamente limitado a la periferia del ámbito al que se dirige, sin alcanzar una justificación válida desde los criterios normativos, técnicos o epistémicos que rigen dicho campo. En cambio, por oposición, un razonamiento práctico centrado (no periférico) es aquel que se realiza con conocimiento o experticia respecto al campo donde recae la acción, considerando criterios normativos, técnicos y epistémicos del mismo. Aunque el ser centrado no quita la posibilidad de error, el núcleo de la cuestión es que el error en el RPP proviene, precisamente, de su carácter de periférico. Si esto lo aplicamos a algún caso de IA, podemos ejemplificarlo del siguiente modo:

*i.* *Premisa mayor (norma general):* toda denuncia falsa de robo con violencia constituye un acto moralmente reprochable y legalmente punible (es un delito).
*ii.* *Premisa menor (identificación del caso):* Veripol es una herramienta de IA diseñada

para detectar denuncias falsas de robos con violencia, mediante el análisis de patrones lingüísticos en los informes policiales (Quijano-Sánchez *et al.*, 2018).

*iii.* *Conclusión:* la Policía Nacional de España implementa Veripol en 2018, bajo la dirección del inspector Miguel Camacho, con el respaldo del Ministerio del Interior (Kolotúshkina, 2018).

El análisis del sistema Veripol ejemplifica de forma clara el fenómeno del RPP. En octubre de 2024, Veripol dejó de estar operativo debido a su carencia de validez judicial y a problemas técnicos significativos en su funcionamiento. Uno de los principales fue la escasa representatividad de la muestra utilizada para su entrenamiento: solo 1122 reportes policiales, de los cuales 534 eran verdaderos y 588 falsos (Martínez, 2024), una cantidad insuficiente si se considera que se reportaban cerca de 30.000 denuncias anuales en España. A esto se sumó la ausencia de expertos legales entre los desarrolladores y la falta de formación adecuada del personal policial encargado de su uso.[1] Esta alfabetización técnica y jurídica mínima debió estar a cargo de profesionales capaces de comprender tanto el funcionamiento del sistema como sus implicancias normativas. En este sentido, Veripol muestra cómo un sistema puede haber sido desarrollado con coherencia interna –en términos de fines y medios–, pero fracasar por falta de racionalidad externa; es decir, por ignorar los conocimientos específicos del campo en el que se implementa. Desde la perspectiva del razonamiento práctico, el caso Veripol constituye un ejemplo concreto del error estructural que caracteriza al RPP: la adopción de decisiones técnicamente estructuradas, pero epistémicamente incompletas frente al dominio real de aplicación.



Si profundizamos, la premisa mayor del razonamiento que condujo a su implementación no es necesariamente falsa. En la premisa menor –la identificación del caso concreto y la competencia técnica para aplicarlo– se presentan los errores que desembocan en una acción inadecuada. De hecho, uno de los errores fue que Veripol no analizaba directamente las denuncias de los ciudadanos, sino lo redactado por los agentes, quienes a su vez no estaban capacitados. Por ello, podemos afirmar que los desarrolladores construyeron Veripol con desconocimiento de aspectos legales esenciales, aun cuando la IA iba a operar en ese ámbito; lo construyeron desde la periferia de la experticia jurídica.

## 3. Alfabetización recíproca bicondicional (ARB) como exigencia de una ética aplicada en el fenómeno de la IA

La dinámica actual del desarrollo y aplicación de la IA revela la presencia de múltiples casos de RPP distribuidos en los distintos sectores implicados. En primer lugar, los académicos tecnocientíficos y expertos de la industria deliberan, diseñan e implementan

[1] Veripol no requería una capacitación compleja, ya que operaba de forma automatizada; sin embargo, el entrenamiento del personal se centró en técnicas de detección de mentiras, lo que generó tensiones con la lógica misma de su implementación.

tecnologías sin atender a las complejas implicancias éticas, sociales y jurídicas propias del campo, lo que constituye un claro caso de RPP desde el ámbito científico e industrial, al actuar fuera del alcance de su experticia. En segundo lugar, académicos de humanidades formulan reflexiones, críticas o propuestas en torno a la IA, pero lo hacen con escasa o nula comprensión técnica del fenómeno, lo que limita la precisión y eficacia de sus intervenciones. Un ejemplo de ello es la crítica radical a la IA basada exclusivamente en ChatGPT, sin diferenciar entre IA generativa e IA predictiva, ni haber utilizado dichas herramientas en la práctica. Este es un caso evidente de RPP desde las humanidades. Finalmente, en el ámbito político y ciudadano, los proyectos de ley sobre IA –y su posterior implementación o aceptación– se basan muchas veces en asesorías provenientes de expertos que también operan desde un RPP, ya sea por falta de formación técnica (en el caso de humanistas) o por desconocimiento normativo y ético (en el caso de técnicos). Este desajuste se transfiere a los políticos y ciudadanos, reproduciendo el RPP a nivel estructural y social.

La ética no es, en su origen, una simple especulación respecto a lo que es bueno o justo, sino que, más bien, busca saber rigurosamente según su objeto qué es bueno y justo, pero también, cómo realizarlo aquí y ahora. De ahí que su virtud esencial consiste en una recta razón en el obrar, conocida comúnmente como *Φρόνησις* (*phronēsis*) o "prudencia". En este sentido, el fin de la ética no es el saber o la teoría, sino la acción u obrar, tal como refiere Aristóteles en la *Ética a Nicómaco* (Aristóteles, 2014). Por ello mismo, el problema del RPP es un problema epistémico, pero también fundamentalmente ético, dado el impacto de las prácticas en el desarrollo y la implementación de la IA en los sujetos implicados y afectados.



Ahora bien, entre los posibles medios de solución o disminución del RPP en los tecnocientíficos y académicos, en este caso, de humanidades, está la alfabetización recíproca bicondicional (ARB). Con alfabetización nos referimos a aquel proceso continuo de aprendizaje (continuo porque implica una constante actualización) que consiste en el conocimiento de los principios, elementos y uso de estos dentro de un campo específico. Esta alfabetización es, además, recíproca porque requiere la actividad de formar tanto por parte de los tecnocientíficos a los humanistas en su campo, como también de los humanistas a los tecnocientíficos en su área. Es, finalmente, bicondicional respecto a la finalidad de esta alfabetización, la cual consiste en alcanzar el bien común de los ciudadanos, lo cual no podrá darse de modo adecuado o pleno si, precisamente, no actúan ambas áreas en conjunto, tanto alfabetizándose mutuamente como mediante la divulgación rigurosa hacia los ciudadanos.

La ARB requiere una base clara que permita identificar que los medios elegidos para la misma y los fines buscados concuerdan con el bien común. Nussbaum propone, a nuestro parecer, una adecuada lista que sirve de base para determinar si, en este caso, algún sistema o alguna herramienta IA que se pretende desarrollar e implementar se ordena o no al bien común propiamente humano. Así, se enumeran diez capacidades que constituirían el bien integral del ser humano y su dignidad de

un modo concreto (Nussbaum, 2003). Estas capacidades cuyo despliegue debe ser garantizado y fomentado son: i) la vida; ii) la salud corporal; iii) la integridad corporal; iv) los sentidos, la imaginación y el pensamiento; v) las emociones; vi) la razón práctica; vii) la afiliación; viii) la relación con otras especies; ix) el juego; y x) el control sobre el propio entorno, ya sea referido a la elección política o a la posesión de propiedades o bienes.

Este enfoque en las capacidades supone que estas condiciones básicas constituyen la calidad de vida, teniendo por ello un valor moral, con la consecuente exigencia social de atender y asegurar tales condiciones. Este enfoque entiende las capacidades en el sentido de la δύναμις (*dýnamis*) aristotélica, la cual puede referirse a facultades o capacidades poseídas de modo innato, pero que igualmente puede referirse al contexto concreto que posibilita el despliegue de aquellas.

Ahora bien, la ARB tiene por finalidad asegurar que, en el desarrollo e implementación de la IA, en sus variadas aplicaciones, se garantice y fomente el despliegue de las capacidades enumeradas anteriormente. Y lo asegura en cuanto que, entre el contexto actual del problema y las soluciones propuestas en materia de IA, se requiere un diagnóstico preciso que no es posible sin una problematización adecuada o, dicho en términos ya usados, una auténtica mirada crítica. En este sentido, la ARB eliminaría el RPP en cuanto supone el trabajo en dos tipos de alineamientos, el primero valórico llamado en la literatura *value alignment* (Gabriel, 2020), mientras que al segundo lo llamaremos, análogamente, "alineamiento técnico".



### 3.1. Alineamiento valórico

El desarrollo tecnológico plantea cada vez con más urgencia este alineamiento como desafío; es decir, cómo diseñar inteligencias artificiales que se ajusten a los valores humanos y a los marcos legales vigentes. Muchas decisiones en la creación de sistemas de IA no son puramente de orden técnico, sino que involucran juicios éticos y jurídicos que los diseñadores técnicos no están necesariamente preparados para asumir. Además, este problema se complejiza por la diversidad de sistemas éticos, culturales y legales entre países. Aunque no es una pregunta técnica en sí, el alineamiento valórico es esencial para garantizar un desarrollo tecnológico seguro en orden a las capacidades humanas. Por ello, la participación activa de académicos humanistas es indispensable en este proceso. Ejemplos que trataremos, como el caso COMPAS, donde los resultados de predicción de reincidencia varían según la noción de imparcialidad empleada, y muestran que estas decisiones afectan directamente a la justicia y a los derechos fundamentales.

### 3.2. Alineamiento técnico

En el contexto actual, los estudios en humanidades requieren un mínimo de alfabetización técnica en IA para analizar adecuadamente fenómenos sociales cada vez más entrelazados con el desarrollo tecnológico. Sin una comprensión

básica de la dimensión técnica, resulta imposible realizar una reflexión amplia y profunda. Esto se evidencia en distintos campos: en sociología, al estudiar fenómenos como la cesantía o la desigualdad empresarial; en derecho, al abordar problemas como la propiedad intelectual o la ambigüedad normativa del GDPR respecto a la explicabilidad; y en ética, donde persiste una visión reduccionista centrada en enfoques consecuencialistas. En este marco, es crucial distinguir entre analfabetismo conceptual (no comprender qué es la IA) y analfabetismo funcional (no saber cómo se usa), pues ambos limitan seriamente el análisis crítico, cayendo muchas veces en RPP. Tales alineamientos, al igual que la ARB, por la misma exigencia de bien común de los ciudadanos a quienes afecta de uno u otro modo el desarrollo e implementación de la IA, tienen también una relación bicondicional. A medida que las técnicas de *machine learning* se perfeccionan y se aplican para responder a preguntas humanas, los dilemas sociales comienzan también a transformarse en dilemas técnicos (Christian, 2020). Por ello es fundamental que los humanistas comprendan el funcionamiento de la IA. Esto implica que el problema de alineamiento es el desafío más importante: lograr que las herramientas de la IA reflejen los valores y las normas de nuestra sociedad para que efectivamente actúen conforme a nuestras intenciones (Christian, 2020).

Si los humanistas no participan del desarrollo tecnológico y los tecnocientíficos desatienden los aportes de las humanidades, la innovación puede generar efectos negativos en la sociedad. A su vez, ignorar las dimensiones técnicas de los cambios sociales y normativos expone a gobiernos, empresas y ciudadanos a riesgos no previstos. En ambos casos, el problema estructural es el RPP. Por ello se vuelve imprescindible fomentar un alineamiento valórico en la tecnología y un alineamiento técnico en las humanidades, a través de la ARB. Parafraseando a Kant respecto al abordaje del fenómeno de la IA desde las humanidades, en concreto desde la ética, los conceptos éticos sin contenido técnico son vacíos y los desarrollos técnicos sin conceptos éticos son ciegos (Kant, 2009).



## 4. Alfabetización en IA en la comunidad académica humanista

Ya considerada la importancia de alinear el desarrollo de la IA a los valores humanos, nos enfocaremos en una de las aristas de la ARB; esto es, la alfabetización en IA en la comunidad académica humanista. A continuación, proponemos una breve guía de los términos y las relaciones que, creemos, componen el conocimiento mínimo para reflexionar sobre las implicaciones de la IA en la vida humana. En la primera parte de esta sección, realizamos una síntesis de las nociones esenciales para hablar de esta tecnología en términos generales. Para ello se abordan los conceptos de inteligencia artificial, aprendizaje por datos, aprendizaje profundo, redes neuronales y opacidad. En la segunda parte, se continúa la distinción de terminología y se distingue entre los dos tipos principales de análisis éticos que estudian la IA: la ética de datos y la ética de modelos. Esta distinción responde a la estructura misma de los sistemas de *machine learning*. Finalmente, en la tercera parte se exploran las

peculiaridades de los dos tipos de modelos que actualmente llamamos inteligencia artificial, identificando sus principales beneficios y obstáculos en la promoción de las personas y la sociedad.

### 4.1. Distinciones esenciales al hablar de inteligencia artificial

Comenzamos aclarando tres conceptos clave que comparten muchas similitudes: IA, *machine learning* y *deep learning*. Primeramente, la IA es un campo de estudio interdisciplinario cuyo propósito es construir entidades artificialmente inteligentes. Hay dos escuelas de pensamiento distintas que se enfocan en este objetivo. La primera busca comprender cómo se representa el conocimiento mediante la lógica y los procesos racionales. Esta teoría tuvo mucho éxito durante la segunda mitad del siglo XX y se basa en programación "tradicional". El segundo método intenta replicar el pensamiento, aprendizaje y comportamiento humanos. Este camino fue elaborado sin mucho éxito a mediados del siglo XX estudiando las redes neuronales cerebrales, pero resurgió con fuerza las últimas dos décadas y ha permitido el desarrollo de muchos de los avances que llamamos IA actualmente (Russell & Norvig, 2021).

Podemos ver, entonces, que el concepto de IA denota la creación de una entidad artificial inteligente, ya sea que entendamos inteligencia como razonamiento lógico o como modo humano de pensar. Ahora bien, este es un objetivo, un propósito. La terminología "inteligencia artificial" no conlleva una metodología concreta para llevar a cabo este propósito. La actual revolución tecnológica está fundada en un avance de la programación, que ha madurado considerablemente las últimas décadas –desde los primeros intentos usando metodologías booleanas manuales, hasta el uso del razonamiento probabilístico y automatización basada en datos (Russell & Norvig, 2021)–, permitiendo los avances en materia de *machine learning* o aprendizaje automatizado. Dado que los sistemas de inteligencia artificial actuales se desarrollan mayormente mediante métodos de *machine learning*, ambos términos –AI y *machine learning*– se suelen usar de forma intercambiable, aunque sean conceptos diferentes.



Las técnicas de *machine learning* varían, pero todas se fundan en dos elementos: un algoritmo de razonamiento probabilístico y una gran cantidad de datos. El algoritmo es la parte que se programa manualmente y donde se diseña lo que la IA puede y no puede hacer. Este algoritmo es entrenado usando extensas bases de datos. Hinton (2024), uno de los padres de la IA, lo explica así: los modelos de *machine learning* se componen de dos partes: una que establece cómo aprende y otra que determina qué aprende. El algoritmo procesa la información de los datos y "aprende" a identificar patrones, ya sea agrupando los elementos (clasificación) o identificando tendencias (regresión). El resultado de todo el proceso es llamado modelo de *machine learning* (**Figura 1**).

**Figura 1. Modelo de *machine learning***

| Componente | Función | Descripción |
|---|---|---|
| Datos | Qué aprende la IA | Grandes cantidades de información en texto e imagen |
| Algoritmo | Cómo aprende la IA | Parte programada a mano que determina de qué forma se procesan los datos |

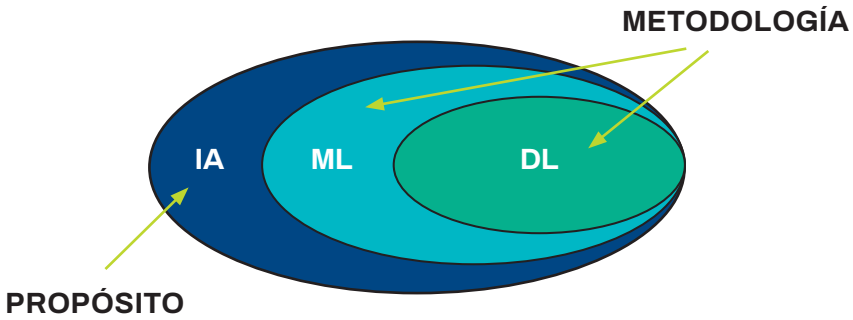
Fuente: elaboración propia.

Los modelos de *machine learning* manifiestan un salto inmenso desde la programación "tradicional" en que todo se programa manualmente. Esto se debe fundamentalmente a dos razones. Primero, al incorporar una fase de entrenamiento a partir de datos, puede generalizar sobre datos desconocidos, lo que permite que el algoritmo mejore su desempeño mediante experiencia (Wilhelm & Zweig, 2024). Segundo, estos sistemas pueden utilizarse para resolver problemas complejos que no sabemos codificar con la programación tradicional (Russell & Norvig, 2021). Estas ventajas han empujado un cambio en el modo de entender el objetivo de crear entidades artificialmente inteligentes. En las décadas en que la programación manual tuvo su auge, se buscaba la creación de sistemas inteligentes mediante el desarrollo de la lógica formal. Durante este tiempo, el término inteligencia se asoció a razonamiento matemático y sin errores. En los últimos años, en cambio, el auge del *machine learning* ha conllevado una transformación en la forma en que se entienden estos agentes artificiales, asociando su inteligencia con la integración de información, la adaptación al medio y la capacidad de resolver problemas. En este sentido, se está aspirando a replicar el comportamiento humano más que a generar razonamiento lógico.



El tercer concepto cuya distinción es clave es *deep learning* o aprendizaje profundo, un tipo de *machine learning* que revolucionó el campo de la IA (**Figura 2**). Esta técnica utiliza múltiples capas de elementos informáticos simples y ajustables al desarrollar el algoritmo de entrenamiento. El término *deep* se refiere a las numerosas capas de procesamiento por las que pasan los datos desde que son ingresados hasta arrojar resultados (Russell & Norvig, 2021). Estos circuitos están organizados imitando las redes neuronales humanas y su sistema de intensidad de estímulos. Es importante recalcar que "redes neuronales" es una analogía biológica; es decir, el modelo funciona como un cerebro humano, sino que la información se procesa en nodos interconectados al modo de neuronas.

**Figura 2. Relación entre inteligencia artificial, *machine learning* y *deep learning***




Fuente: elaboración propia.

Los modelos de *deep learning* aportan metodologías potentes para avanzar en el proyecto de desarrollar agentes artificiales y presentan dos ventajas principales. En primer lugar, dado que los datos de entrada se procesan mucho más que con métodos más simples de *machine learning*, sus resultados son más precisos (Maclure, 2021). En segundo lugar, soluciona el problemático dilema entre expresividad y complejidad que se produce en la programación manual: cuanto más extenso es el código, mayor es la expresividad del modelo, pero más difícil es para un computador procesarlo. Por el contrario, cuanto más computable es el código, menos expresivo es. Con las redes neuronales profundas, las representaciones no son simples, pero sí fácilmente computables con el hardware adecuado (Russell & Norvig, 2021). En este sentido, los modelos de *deep learning* desempeñan hoy en día un papel similar al de los transistores en el siglo XX, permitiendo el desarrollo de tecnologías complejas minimizando recursos. Aun así, es importante destacar que esta economía de recursos es comparativamente menor, pero igualmente enorme. Los modelos de IA actuales requieren enormes cantidades de recursos naturales para funcionar. Los *data center* donde se procesan sus datos usan, además de mucha electricidad, grandes cantidades de agua dulce para mantener la temperatura estable. Como referencia, los edificios de *data center* que procesan las transacciones de una sola criptomoneda (*bitcoin*) gastan más electricidad que países como Chile o Dinamarca (Narayanan & Kapoor, 2024).



Además del gasto de recursos naturales, la desventaja más conocida de los modelos de aprendizaje profundo es opacidad epistémica de los modelos computacionales: debido al gran número de capas que utilizan y a la gran cantidad de datos que gestionan, una vez finalizado el entrenamiento, el algoritmo se convierte en un laberinto. Aunque, en teoría, toda esa información está disponible para su comprobación, resulta casi imposible rastrear cómo cambian los pesos de las interconexiones y los patrones, ya que el modelo es demasiado grande para ser observado exhaustivamente (Barredo *et al.*, 2020). Ni siquiera los propios desarrolladores comprenden completamente cómo

el modelo determina un resultado. De forma similar a contar estrellas, el programador puede examinar el algoritmo, pero en la práctica no puede identificar todos los detalles ni, menos aún, asociar la información y las relaciones relevantes. Si bien el código es accesible *per se*, es inaccesible *per nos*; es decir, la forma en que un modelo de *deep learning* asigna una etiqueta a un punto de datos a menudo es incomprensible para un ser humano (Wilhelm & Zweig, 2024). Dada su opacidad, estos sistemas han sido denominados modelos de caja negra, en contraste con los modelos de caja blanca, fácilmente comprensibles.

Ahora bien, el problema de la opacidad no se remite solo a los modelos de redes neuronales, sino que también se da en otros modelos de *machine learning*. Dado que la opacidad se debe a la cantidad de capas y elementos del algoritmo, esta no es exclusiva del *deep learning* de redes neuronales. Otros modelos de *machine learning* también son opacos por su gran cantidad de componentes, como los modelos de regresión logística, análisis de clúster, *support vector machine* (SVM), *random forests*, entre muchos otros. Algunos de estos modelos se basan en arquitecturas interpretables, pero por su tamaño se hacen opacos, como el caso de modelos de *random forest*, cuyo algoritmo combina múltiples árboles de decisión.

Ante la opacidad algorítmica, una solución que ha ganado popularidad recientemente es el uso de IA explicable o *explainable IA* (XAI). Estos modelos de IA más simples e intrínsecamente interpretables se usan como herramienta externa y a posteriori para obtener información sobre el "razonamiento" de una IA opaca mediante simplificación o sensibilidad de categorías (Barredo *et al.*, 2020). Las técnicas de simplificación conectan las conclusiones de la IA opaca con los datos proporcionados inicialmente y así identifican las categorías que tuvieron más importancia en el modelo de la IA opaca. Dos ejemplos conocidos son LIME (Ribeiro *et al.*, 2016) que usa funciones lineales y SHAP (Lundberg & Lee, 2017) que usa teoría de juegos para relacionar datos iniciales y finales. Por otro lado, las técnicas de sensibilidad modifican los pesos asignados a cada nodo (incremento o disminución) advirtiendo si se generan cambios en los resultados del modelo. Así, determina la relevancia de la categoría según lo sensible que es el modelo a ella (Barredo *et al.*, 2020).



Estas técnicas de explicabilidad generan una nueva capa epistémica, ya que no dan cuenta del fenómeno a evaluar sino de la forma de evaluarlo que tiene la IA opaca. Por ejemplo, la visualización XAI ayuda a identificar los elementos que utiliza un vehículo autónomo para determinar la ruta; sin embargo, dicha visualización no constituye la lógica de la IA en sí, sino una representación probabilística. En ese sentido, siendo una interpretación de la interpretación, las XAI proporcionan conocimiento de segundo orden que puede generar disociación epistémica (Fleisher, 2022). En paralelo, se han propuesto desde la academia metodologías de evaluación de las explicaciones y marcos de transparencia (Mittelstadt, 2022).

Si bien la opacidad técnica que se genera por el uso de algoritmos tan complejos es altamente problemática, se habla poco de otro tipo que es igual o incluso más

problemática: la opacidad legal. La mayoría de los modelos de IA son producidos por empresas privadas que patentan sus herramientas, por lo que la forma de recolectar y procesar datos queda oculta bajo secreto comercial. Aunque por sí mismo el secreto comercial fomenta la innovación privada, trae problemas a la hora de evaluar si los sistemas de IA desarrollados son seguros. No solo los programadores se enfrentan a la caja negra del aprendizaje profundo, sino que las agencias reguladoras, los usuarios y las personas afectadas no tienen cómo saber la manera en que estas herramientas funcionan, dado que el algoritmo es secreto. Hay empresas que abusan de esto y promocionan sus herramientas de IA con descripciones exageradas de sus resultados e incluso directamente engañando a los compradores, lo que se ha denominado *AI washing*. Uno de los casos más escandalosos fue el de la tecnología Just Walk Out de Amazon. El sistema de reconocimiento facial en las cámaras de los supermercados Amazon Fresh permitió que los usuarios pudieran hacer pago automático de sus productos sin pasar por caja. Esta IA fue luego puesta a la venta por Amazon a múltiples empresas haciendo énfasis en su efectividad. Sin embargo, tras años de comercialización, en 2024 se descubrió que el reconocimiento facial no era realizado por una IA, sino por cientos de trabajadores en la India que memorizaban la fisonomía facial de los compradores.

Adicionalmente, una problemática discutida de los sistemas de *machine learning* es su funcionamiento autónomo, sobre todo en relación con la brecha de responsabilidad o *responsibility gap*, fenómeno planteado por primera vez por Matthias (2004) que 20 años después sigue siendo fuente de numerosas discusiones éticas y de legislación. Considerando la capacidad de la IA para aprender nuevos patrones, un resultado puede no ser previsto por el fabricante ni por el operador. Por ello, cuando se producen efectos indeseables tras las sugerencias de la IA, a veces es imposible atribuir responsabilidad moral ni legal a los actores involucrados (usuarios, programadores, empresas tecnológicas y otros), lo que ha sido especialmente preocupante en contextos clínicos (Santoni de Sio & Mecacci, 2021) y de vehículos de conducción autónoma (Nyholm, 2023).



### 4.2. Consideración de los componentes de la IA

Parte de la alfabetización en IA necesaria para los académicos humanistas es distinguir los desafíos de la IA en cuanto algoritmo y en cuando procesamiento de información. Siguiendo la composición de las herramientas de *machine learning*, la reflexión sobre IA se puede subdividir en dos partes: el estudio del manejo de la información utilizada y el estudio de la gestión del algoritmo. A lo primero se le suele llamar ética de datos, mientras que lo segundo lo llamaremos análogamente ética de modelos.

A diferencia de otras tecnologías, los modelos de *machine learning* no solo reflejan la realidad, sino que la interpretan y reproducen. Un sistema de IA observa datos, construye un modelo del mundo y lo utiliza como hipótesis para resolver problemas (Russell & Norvig, 2021). Por esto, la IA funciona a la vez como asistente y como espejo de la sociedad, amplificando patrones culturales profundamente arraigados que

muchas veces pasan desapercibidos. Algunos de estos patrones pueden ser peligrosos o no deseados, como la exclusión por religión, raza, sexo y otros, por lo que los datos utilizados y la arquitectura del algoritmo requieren una revisión ética rigurosa.

*4.2.1. Manejo de información (ética de datos)*

Desde el punto de vista de los datos usados para el entrenamiento de un sistema de IA, hay varios desafíos que considerar. Estos se pueden agrupar en adquisición y sistematización, en otras palabras, cómo se recolecta la información y cómo se la prepara para ser usada.

El primer desafío de adquisición de información es simple conceptualmente, pero sus repercusiones son inmensas: cómo las empresas desarrolladoras de IA obtienen la data para entrenar sus productos. La discusión se centra en la propiedad de la data y en el consentimiento de su uso para fines de entrenamiento, distinguiendo cuáles los datos personales son privados y cuáles libres y, sobre ello, si empresas desarrolladoras debieran pagar a los dueños incluso por el uso de datos libres, discusión que tiene casi 30 años (Nissenbaum, 1998). Ligado a la propiedad de los datos está el problema del consentimiento de usuarios individuales sobre el uso de su información y para qué fines. Por un lado, el *data mining* puede ofrecer beneficios como comprender mejor a personas similares (Fairfield y Engel, 2015), pero por otro lado plantea riesgos de abuso de poder, como el procesamiento de información personal para vigilancia (Zuboff, 2015). En los últimos años, esta discusión se ha extendido a la propiedad creativa en casos de entrenamiento de IA generativa, pues estos modelos son capaces de crear textos e imágenes excepcionales, pero gracias a que fueron entrenadas con material producido por seres humanos. En esta misma línea, el *New York Times* demandó a OpenIA por utilizar su material sin autorización y una maniobra similar se espera de Studio Ghibli, cuya estética está siendo replicada sin consentimiento por millones de personas.



El segundo desafío de la ética de datos se relaciona con el preprocesamiento de los mismos: en concreto, cómo se ordenan y limpian los datos antes del entrenamiento del modelo. Este proceso implica tomar decisiones fundamentales que afectan directamente los resultados del sistema, como la elección de categorías, la inclusión o exclusión de variables sensibles, y el uso de técnicas para completar información faltante. Primeramente, al momento de preparar la data para entrenar un algoritmo de *machine learning*, se necesita definir y estandarizar los criterios con que caracterizar las personas, cosas o fenómenos a considerar. La forma de ordenar los datos va a tener un impacto directo en los resultados. Además, en caso de que el algoritmo trate sobre fenómenos humanos, se debe definir qué conceptos personales distinguir o unir (como sexo y género) y cuáles excluir (como religión, raza o categorías protegidas de un país), evaluando la pertinencia y legalidad de estos datos.

Otro aspecto a considerar es el tratamiento de bases de datos incompletas. Para mejorar la precisión de los modelos, a menudo se emplean técnicas de *data augmentation* que buscan completar la información faltante mediante inferencias

automáticas o generación de datos sintéticos. A pesar de la utilidad de estas técnicas, pueden reducir la precisión del modelo, introducir sesgos adicionales y generar información artificialmente coherente, pero no necesariamente verdadera (Hao *et al.*, 2024). Asimismo, cuando se tratan elementos difíciles de cuantificar, se puede usar indicadores indirectos como los *proxies* (sustitutos) o la operacionalización. Estas técnicas facilitan el análisis de fenómenos humanos, pero pueden llevar a la confusión del fenómeno en cuestión y su metodología de medición. Por ejemplo, a la hora de medir el riesgo de reincidencia se puede usar la cantidad de arrestos para cuantificar los delitos –siendo que estas acciones no son exactamente lo mismo–, sustituyendo un elemento desconocido por uno conocido (Christian, 2020). En otro caso, se puede determinar la calidad del profesorado operacionalizando "buen profesor" en términos de las notas de sus alumnos, ranking del colegio y otros elementos cuantificables (Biddle, 2022). Al revisar un modelo, es fundamental tener en cuenta qué *proxies* u operacionalizaciones se usaron en su diseño para no inferir conclusiones erróneas. También es importante tomar en cuenta que la IA también puede generar *proxies* inadvertidamente como, por ejemplo, determinar la probabilidad de desarrollar cáncer de piel solo considerando el país de nacimiento de una persona, usando el país como *proxy* de cantidad de exposición a rayos UV.

Adicionalmente, es importante destacar que no todas las IA son entrenadas con datos etiquetados, por los que estos desafíos no se dan por igual en todos los casos. Algunos modelos incorporan información que está vinculada a una etiqueta, lo que se llama aprendizaje supervisado. Por ejemplo, una base de datos de imágenes de animales puede tener asociada una o más etiquetas por elemento ("mamífero", "gato", "perro", "galgo"). En cambio, otros modelos de procesamiento usan bases de datos sin etiquetas asignadas, por lo que la IA genera sus propias formas de agrupar los elementos desde cero. Estos modelos generan sus propias categorías, que luego pueden ser ratificadas y objetadas (aprendizaje no supervisado con y sin reforzamiento). En casos de aprendizaje no supervisado, no se presentan varios de los desafíos de sistematización de los datos por lo que este tipo de modelos puede resultar más neutral para ciertos proyectos. Sin embargo, conllevan desafíos igualmente, como el reforzamiento de sesgos históricos o el uso inadvertido de *proxies* a la hora de identificar patrones.



*4.2.2. Empleo del algoritmo (ética de modelos de IA)*

Además del tratamiento de datos, la arquitectura del modelo plantea repercusiones éticas que requieren análisis. Dentro del desarrollo del modelo se pueden distinguir tres etapas –entrenamiento, validación e implementación–, cada una con sus desafíos.

Durante la etapa de entrenamiento se definen elementos cruciales como el tipo de algoritmo y los hiperparámetros. La decisión de usar cierto tipo de algoritmo de *machine learning* acarrea sus propias ventajas y desventajas, como la atingencia y precisión para representar un fenómeno, su transparencia (tanto técnica como legal) y su computabilidad. Asimismo, la selección y el ajuste de hiperparámentros condiciona la forma en que la IA aprende, como en la selección del número de capas y nodos de una

red neuronal o la cantidad de ramificaciones que tendrá un árbol de decisión. Al igual que en la elección del algoritmo, estas variables reflejan los deseos de los programadores y están sujetas a la evaluación ética si la IA afecta la vida de las personas.

Un dilema clave en esta etapa es cómo lograr imparcialidad cuando los datos históricos están sesgados; es decir, donde hay identificaciones tendenciosas o que simplemente ya no son correctas, como por ejemplo que los hombres son médicos y las mujeres enfermeras. Frente a esto, la literatura propone varias formas para lograr imparcialidad algorítmica como la calibración del modelo (misma capacidad predictiva para los grupos) y ecualización de probabilidades (nivelación de la tasa de falsos positivos y negativos de los grupos). Lamentablemente, la satisfacción de uno de estos criterios compromete al otro y es imposible satisfacer ambas métricas simultáneamente (Corbett-Davies *et al.*, 2023), por lo que se requiere reflexión legal y ética sobre qué tipo de imparcialidad se prefiere en diferentes contextos. Un caso especialmente delicado relacionado con esta problemática es el sistema COMPAS, utilizado en diversos estados de Estados Unidos durante los últimos 25 años para evaluar el riesgo de reincidencia penal. Esta IA caracteriza las personas negras como más propensas a la reincidencia, negándoseles libertad condicional porque históricamente ha habido más arrestos de gente negra que blanca. Por otro lado, numerosas personas blancas han sido falsamente consideradas de baja peligrosidad y vuelto a delinquir bajo libertad condicional (Biddle, 2022). Los desarrolladores de COMPAS sostienen que han hecho todo lo posible por hacer que el algoritmo haga predicciones equitativas por medio de calibración, mientras que otras organizaciones alegan que sus tasas de falsos negativos y falsos positivos son inadmisibles (Christian, 2020).



Una vez completado el entrenamiento del modelo, se realiza su verificación y validación para que el modelo esté bien empleado y que sea eficaz en representar el fenómeno deseado. Durante esta fase se comprueba que el modelo funcione correctamente según los criterios que se consideren pertinentes. Esos criterios varían según los objetivos técnicos, pero también criterios éticos, culturales y legales, por lo que es importante transparentarlos en lo posible. Además, es importante que en esta etapa se use un conjunto de datos nunca procesado por el algoritmo, de modo que se compruebe la fiabilidad de su funcionamiento y se corrijan sus deficiencias. Finalmente, en la etapa de implementación del modelo inicia con un testeo en el mundo real para asegurarse de que el modelo funciona fuera de las constreñidas condiciones de laboratorio. Aquí muchos modelos fallan de formas inesperadas. Un caso emblemático de este tipo de desafíos es el de Google en 2015 cuando lanzó su sistema de reconocimiento de imagen. La data de entrenamiento y validación contenía en su mayoría personas blancas, por lo que, al testearlo en el mundo real, resultó catalogar a las personas negras como gorilas. Tres años después, este problema seguía siendo imposible de solucionar, por lo que Google simplemente quitó la etiqueta “gorila” de las respuestas.

Otro elemento importante de revisar a la hora de la implementación es cómo se transmiten los resultados obtenidos, sobre todo en casos en que la IA ayuda en la toma de decisiones. Cuando se quiere representar la eficacia de los resultados de una

IA –sobre todo en modelos de reconocimiento de imagen–, se muestra el porcentaje de acierto en cierta cantidad de intentos. Por ejemplo, una IA que identifica cáncer en las imágenes de una resonancia magnética puede tener 90% de efectividad, pero es bastante diferente si eso refiere a *top-2 90%* (90% en dos intentos) o *top-5 90%* (90% en cinco intentos) (Narayanan & Kapoor, 2024). Es importante estar al tanto de este tipo de expresiones porque pueden ser engañosas para el usuario y provocar un exceso de confianza que favorece estafas y falta de adecuada revisión de los resultados.

### 4.3. Repercusiones de distintos tipos de IA

Un último elemento esencial para llevar a cabo una reflexión atingente y significativa de la inteligencia artificial es la diferenciación entre IA predictiva e IA generativa. La IA predictiva, que ha sido ampliamente utilizada en las últimas dos décadas, se centra en el análisis de datos históricos para identificar patrones y proyectarlos hacia situaciones nuevas. Por su lado, la IA generativa es la reciente protagonista con herramientas como ChatGPT, Gemini o DALL-E, que producen contenido nuevo –como texto o imágenes–, a partir de instrucciones del usuario, imitando material de internet. Aunque ambos tipos de herramientas se fundamentan en modelos de *machine learning* y grandes volúmenes de datos, sus propósitos, beneficios, riesgos y aplicaciones difieren de forma significativa, por lo que deben ser diferenciadas al momento de evaluar sus implicaciones en la sociedad.



La IA predictiva se desarrolló desde el análisis de *big data* y ha tenido un avance paulatino pero constante. En diversos contextos, especialmente empresariales y gubernamentales, ha mostrado su utilidad en la asistencia de tareas repetitivas, peligrosas o difíciles. Puede automatizar operaciones como la moderación de contenido en redes sociales, evitando que seres humanos se expongan a material violento, como asesinatos o pornografía infantil (Narayanan & Kapoor, 2024). Más aún, dada su capacidad de anticipar comportamientos futuros, es un apoyo en la toma de decisiones como diagnósticos médicos, crediticios, laborales y judiciales, entre otros, actuando como sistemas de toma de decisión automática (o ADM, por sus siglas en inglés). Por ejemplo, puede servir para aconsejar en situaciones en que falta un especialista, como está sucediendo en el Reino Unido, donde, frente a la escasez de radiólogos infantiles, se están desarrollando herramientas de IA predictiva para asesorar a profesionales no especializados en anatomía pediátrica (Shelmerdine, 2024).

A pesar de las numerosas ventajas, guarda una limitación importante: solo puede predecir en la medida en que el futuro se parezca al pasado. Al entrenarse con datos históricos, estos sistemas adquieren la capacidad de evaluar información nunca vista, siempre y cuando esta nueva información tenga cierta similitud con los datos de entrenamiento. De ahí que su rendimiento disminuya cuando intenta abordar fenómenos sociales cambiantes o situaciones para las que no hay antecedentes suficientes. Esto se condice con el recelo de los últimos años en la automatización de decisiones. Los médicos temen dar un diagnóstico erróneo (Santoni de Sio & Mecacci, 2021), los jueces están descontentos con las sugerencias de libertad condicional

(Biddle, 2022), los ciudadanos temen ser discriminados cuando un banco rechaza su solicitud de préstamo (Vredenburgh, 2021). Lo anterior se vuelve más preocupante considerando la opacidad algorítmica y el secreto comercial que dificultan verificar el funcionamiento de la IA. Frente a esto, muchos países han imitado el reglamento general de protección de datos de la Unión Europea, regulando las explicaciones de los sistemas de ADM en pos de lo que se ha llamado "el derecho a la explicación" (Goodman & Flaxman, 2017).

Por otra parte, la IA generativa (o GenAI, por sus siglas en inglés), además de identificar patrones, produce contenido en formato texto, imagen, música e incluso código imitando el estilo de los ejemplos con los que fue entrenada. Estos sistemas se entrenan con grandes volúmenes de datos extraídos de Internet y generan respuestas probabilísticamente plausibles. Aunque también pueden ser usadas a nivel de grandes instituciones, su gran auge ha sido entre el público general, lo que ha desencadenado además el especial interés y preocupación por este tipo de IA en la prensa y redes sociales.

La generación de texto se basa en un sistema similar –guardando las proporciones– a las sugerencias de frases al escribir un correo electrónico. El modelo analiza lo escrito por el usuario, predice qué palabras son más probables en ese contexto y las concatena para formar párrafos coherentes (Narayanan & Kapoor, 2024). Esta generación no se basa en la comprensión del asunto, sino en asociaciones estadísticas extraídas de material disponible en la web, por lo que no ofrece conocimiento ni juicio en la materia. Uno de los grandes riesgos de estos modelos es que a menudo se generan secuencias de palabras con sentido, pero con información falsa, fenómeno conocido como "alucinación". Junto a ello, es importante considerar que estos sistemas no consideran conocimientos no digitalizados como la cultura oral y que asimilan la idea de mundo y sesgos culturales del idioma en que entrenan. Por estas razones, este tipo de IA no es adecuado para asistir en la toma de decisiones, especialmente en casos donde se requiere precisión y se afecta la vida de las personas.



Los sistemas de generación de imagen, en cambio, utilizan otras metodologías. Usualmente emplean dos redes neuronales que actúan como adversarios (GAN): una crea imágenes a partir de ruido visual y la otra las evalúa según su grado de similitud con imágenes reales. El proceso iterativo permite generar resultados similares a los que se encuentran en internet, y añadir elementos y ajustes según las preferencias del usuario, lo que permite crear fácilmente ilustraciones y fotos falsas (*deepfake*) como el viral caso del Papa Francisco vistiendo una inmensa chaqueta Balenciaga. Estas capacidades pueden utilizarse con fines maliciosos, como el robo de identidad o la creación de contenido sexual con los rostros de menores de edad, que incorpora peligros sociales, psicológicos y legales.

## Conclusión

En las disciplinas humanistas de la academia, la alfabetización en IA se ha vuelto una exigencia ética y epistémica. Reflexionar sobre los impactos sociales, políticos y jurídicos de la IA sin comprender su funcionamiento técnico genera análisis errados, poco pertinentes e incluso contraproducentes, fruto del problema estructural base que denominamos razonamiento práctico periférico (RPP). Confundir los componentes y las funciones de la IA –por ejemplo, no distinguir entre IA generativa e IA predictiva, o entre dilemas asociados a los datos (como sesgos históricos) y aquellos vinculados a los modelos (como el tipo de algoritmo usado)– puede llevar a decisiones mal informadas, implementación inadecuada de tecnologías y vulneración de derechos fundamentales. Más aún, la falta de comprensión técnica en los análisis humanistas no solo empobrece el debate, sino que favorece una mirada periférica que impide abordar con profundidad los desafíos éticos y sociales de la IA en nuestro tiempo. Así, el desarrollo acelerado de sistemas de IA exige que los estudios humanistas superen el analfabetismo conceptual y funcional, ya que ningún análisis sobre inteligencia, agencia o autocomprensión de la IA, ni sobre sus implicaciones éticas, sociales o legales, será riguroso si no parte de una comprensión básica del fenómeno técnico. La alfabetización recíproca bicondicional (ARB) ofrece una vía para resolver esta brecha, promoviendo un proceso mutuo y continuo de formación entre humanistas y técnicos. Esto permite a los académicos humanistas desarrollar una mirada crítica verdaderamente informada, fomentando el alineamiento técnico; es decir, la capacidad de entender con precisión los desafíos que plantea la IA. A su vez, esta alfabetización permite que los desarrollos técnicos integren principios éticos y marcos jurídicos sólidos (alineamiento valórico), evitando que los juicios normativos queden al margen del diseño tecnológico. Dado que la falta de diálogo y comprensión mutua entre disciplinas impide que la IA se oriente hacia el bien común, la ARB no solo enriquece el debate, sino que constituye un requisito para una innovación verdaderamente responsable.



## Reconocimiento de uso de IA

Los autores de este artículo declaran que el título y el resumen en lengua portuguesa de este artículo fueron elaborados con el apoyo de una herramienta web. La versión resultante fue igualmente revisada por los autores, quienes asumen plena responsabilidad por su contenido y fidelidad respecto de la versión original en español.

## Bibliografía

Aristóteles (1988). Política. Madrid: Gredos.

Aristóteles (2014). Ética a Nicómaco. Madrid: Gredos.

Barredo, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R. & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. Recuperado de: https://arxiv.org/abs/1910.10045.

Biddle, J. (2022). On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning. Canadian Journal of Philosophy, 52(3), 321-341. Recuperado de: https://philpapers.org/rec/BIDOPR.

Brockman, J. (1995). The Third Culture: Beyond the Scientific Revolution. Nueva York: Touchstone.

Christian, B. (2020). The Alignment Problem: How Can Artificial Intelligence Learn Human Values? Londres: Atlantic Books.

Corbett-Davies, S., Gaebler, J., Nilforoshan, H., Shroff, R. & Goel, S. (2023). The Measure and Mismeasure of Fairness. Journal of Machine Learning Research, 24. DOI: https://dl.acm.org/doi/10.5555/3648699.3649011.

Fairfield, J. & Engel, C. (2017). Privacy as a Public Good. En R. Miller (ed.), Privacy and Power: A Transatlantic Dialogue in the Shadow of the NSA-Affair (95-128). Cambridge: Cambridge University Press. Recuperado de: https://www.cambridge.org/core/books/abs/privacy-and-power/privacy-as-a-public-good/70237D89C5F6238BFCF5F572EB21833E.

Fleisher, W. (2022). Understanding, Idealization, and Explainable AI. Episteme, 19(4), 534-560. Recuperado de: https://philpapers.org/rec/FLEUIA.

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. Minds & Machines, 30, 411-437. Recuperado de: https://arxiv.org/abs/2001.09768.

Goodman, B. & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation”. AI Magazine, 38(3), 50-57. DOI: https://onlinelibrary.wiley.com/doi/epdf/10.1609/aimag.v38i3.2741.

Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, Zhou Z., C. & Tang, H. (2024). Synthetic data in AI: Challenges, applications, and ethical implications. arXiv preprint. Recuperado de: https://doi.org/10.48550/arXiv.2401.01629.

Hinton, G. (2024). The politics and philosophy of AI | LSE Event [video]. Recuperado de: www.youtube.com/watch?v=5Oqbg72xivw.

Jauregui-Volpe, M. (2025). OpenAI Partners with AAU Members and Other Research Institutions to Find New Applications for AI. Association of American Universities, 7 de

marzo. Recuperado de: https://www.aau.edu/newsroom/leading-research-universities-report/openai-partners-aau-members-and-other-research.

Kant, I. (2009). Crítica de la Razón Pura. Ciudad de México: Fondo de Cultura Económica.

Lucas, A., Revell, A. & Davis, K. A. (2024). Artificial intelligence in epilepsy - applications and pathways to the clinic. Nature reviews. Neurology, 20(6), 319-336. Recuperado de: https://pubmed.ncbi.nlm.nih.gov/38720105/.

Lundberg, S. & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. En Actas de la 31ava Conferencia NIPS'17, International Conference on Neural Information Processing Systems (4768–4777). Nueva York Curran Associates Inc. Recuperado de: https://www.researchgate.net/publication/317062430_A_Unified_Approach_to_Interpreting_Model_Predictions.

Maclure, J. (2021). AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind. Minds and Machines, 31(3), 421-438. Recuperado de: https://link.springer.com/article/10.1007/s11023-021-09570-x.

Martínez, L., Boix, A., Briz A., Flores, F., García A., Molina, M, Montes, F., Palma, A., Peris A. & Soriano, A. (2024). Three predictive policing approaches in Spain: VioGén, RisCanvi and VeriPol. Assessment from a human rights perspective. Universidad de Valencia. Recuperado de: www.regulation.blogs.uv.es/files/2024/05/Three-predictive-policing-perspectives-web-17.06.24.pdf.



Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6(3), 175-183. Recuperado de: https://link.springer.com/article/10.1007/s10676-004-3422-1

McKinney, S., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C., King, D. & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94. Recuperado de: https://www.nature.com/articles/s41586-019-1799-6.

Mittelstadt, B. (2022). Interpretability and transparency in artificial intelligence. En C. Véliz (Ed.), The Oxford Handbook of Digital Ethics (378-410). Oxford: Oxford University Press.

Narayanan, A. y Kapoor, S. (2024). AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference. Princeton: Princeton University Press.

Nissenbaum, H. (1998). Protecting privacy in an information age: The problem of privacy in public. Law and Philosophy, 17(5-6), 559-596. Recuperado de: https://link.springer.com/article/10.1023/A:1006184504201.

Nussbaum, M. (2003). Capabilities as Fundamental Entitlements: Sen and Social Justice. Feminist Economics, 9(2-3), 33-59. Recuperado de: https://philpapers.org/rec/NUSCAF.

Nyholm, S. (2023). Responsibility Gaps, Value Alignment, and Meaningful Human Control over Artificial Intelligence. En A. Placani y S. Broadhead (Eds.), Risk and Responsibility in Context (191-213). Nueva York: Routledge. Recuperado de: https://www.researchgate.net/publication/373325103_Responsibility_Gaps_Value_Alignment_and_Meaningful_Human_Control_over_Artificial_Intelligence.

Popenici, S. (2023). The critique of AI as a foundation for judicious use in higher education. Journal of Applied Learning and Teaching, 6(2), 378-384. Recuperado de: https://www.researchgate.net/publication/372194610_The_critique_of_AI_as_a_foundation_for_judicious_use_in_higher_education.

Quijano-Sanchez, L., Liberatore, F., Camacho-Collados, J. & Camacho-Collados, M. (2018). Applying automatic text-based detection of deceptive language to police reports: extracting behavioral patterns from a multi-step classification model to understand how we lie to the Police. Knowledge-Based Systems, 149, 155-168. Recuperado de: https://www.sciencedirect.com/science/article/abs/pii/S095070511830128X.

Ribeiro, M., Singh, S. & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. En Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (97-101). Recuperado de: https://arxiv.org/abs/1602.04938.



Russell, S. & Norvig, P. (2021). Artificial Intelligence: A Modern Approach. Londres: Pearson.

Santoni de Sio, F. & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. Philosophy and Technology, 34(4), 1057-1084. Recuperado de: https://link.springer.com/article/10.1007/s13347-021-00450-x.

Shelmerdine, S. C. (2024). Revolutionising paediatric radiology: the future impact of artificial intelligence. European Radiology, 34, 2294–2296. Recuperado de: https://link.springer.com/article/10.1007/s00330-023-10288-w.

Snow, C. P. (1961). The two cultures and the scientific Revolution. Nueva York: Cambridge University Press.

Trujillo, J. & Vallejo, X. (2007). Silogismo teórico, razonamiento práctico y raciocinio retórico-dialéctico. Praxis Filosófica, (24), 79-114. Recuperado de: https://praxisfilosofica.univalle.edu.co/index.php/praxis/article/view/3135.

Vigo, A. (2013). La conciencia errónea: De Sócrates a Tomás de Aquino. Signos filosóficos, 15(29), 9-37. Recuperado de: https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-13242013000100001.

Vredenburgh, K. (2021). The Right to Explanation. Journal of Political Philosophy, 30(2), 209-229. Recuperado de: https://onlinelibrary.wiley.com/doi/abs/10.1111/jopp.12262.

Whewell, W. (1967). The philosophy of the inductive sciences, founded upon their history. Nueva York: Johnson Reprint.

Wilhelm, A. & Zweig, K. (2024). Hacking a surrogate model approach to XAI. Recuperado de: https://web3.arxiv.org/pdf/2406.16626.

Yáñez, E. (2019). La política en la actualidad: ¿más cerca de la virtud o del vicio? Sapientia, 70(236), 143-154. Recuperado de: https://e-revistas.uca.edu.ar/index.php/SAP/article/view/1837.

Zuboff, S. (2015). Big Other: Surveillance Capitalism and the Prospects of an Information Civilization. Journal of Information Technology, 30, 75–89. Recuperado de: https://journals.sagepub.com/doi/10.1057/jit.2015.5.

# El papel de la Santa Sede en el desarrollo de la ética de la inteligencia artificial

# O papel da Santa Sé no desenvolvimento da ética da inteligência artificial

# *The Role of the Holy See in the Development of Artificial Intelligence Ethics*

**Andrea Ciucci** *



A diferencia de otras innovaciones tecnológicas, el rápido desarrollo de los sistemas de inteligencia artificial ha generado un debate, que no puede ser definido como trivial, sobre la ética que debe caracterizar esta transformación. En los últimos cinco años, la Santa Sede ha contribuido de manera significativa al nacimiento de la "algorética", tanto a través de varias intervenciones del Pontífice como gracias a iniciativas y documentos propuestos por algunos organismos del Vaticano directamente involucrados en este campo. A través de una metodología que ha privilegiado la participación de todos los actores de este proceso, la cuestión de la ética de la inteligencia artificial se ha basado en un juicio positivo, pero no ingenuo, de estas tecnologías. El resultado es una reflexión ética que no se limita a la reducción de posibles riesgos y que, más bien, plantea la pregunta de qué futuro para el hombre y la sociedad se quiere y se puede construir, gracias a la introducción de los sistemas de inteligencia artificial, y cuáles son las condiciones para que esto pueda suceder.

**Palabras clave:** Santa Sede**;** algorética**;** inteligencia artificial; ética**;** Papa Francisco

* Coordinador de la Secretaría de la Academia Pontificia para la Vida, Secretario General de la Fundación RenAIssance para la Ética de la Inteligencia Artificial, Ciudad del Vaticano. Correo electrónico: andrea.ciucci@pav.va. ORCID: https://orcid.org/0009-0006-5859-3023.

Ao contrário de outras inovações tecnológicas, o rápido desenvolvimento dos sistemas de inteligência artificial gerou um debate, que não pode ser definido como trivial, sobre a ética que deve caracterizar essa transformação. Nos últimos cinco anos, a Santa Sé contribuiu significativamente para o nascimento da "algorética", tanto por meio de várias intervenções do Pontífice quanto graças a iniciativas e documentos propostos por alguns organismos do Vaticano diretamente envolvidos neste campo. Através de uma metodologia que privilegiou a participação de todos os atores desse processo, a questão da ética da inteligência artificial baseou-se em um julgamento positivo, mas não ingênuo, dessas tecnologias. O resultado é uma reflexão ética que não se limita à redução de possíveis riscos e que, ao contrário, levanta a questão de qual futuro para o homem e a sociedade se deseja e se pode construir, graças à introdução dos sistemas de inteligência artificial, e quais são as condições para que isso possa acontecer.

**Palavras-chave:** Santa Sé; algorética; inteligência artificial; ética; Papa Francisco

*Unlike other technological innovations, the rapid development of artificial intelligence systems has sparked a debate, which cannot be defined as trivial, on the ethics that should characterize this transformation. Over the past five years, the Holy See has contributed significantly to the birth of "algorethics", both through various interventions by the Pontiff and thanks to initiatives and documents proposed by some Vatican bodies directly involved in this field. Through a methodology that has privileged the participation of all actors in this process, the question of the ethics of artificial intelligence has been based on a positive, but not naive, assessment of these technologies. The result is an ethical reflection that is not limited to reducing potential risks but rather raises the question of what future for humanity and society we want and can build thanks to the introduction of artificial intelligence systems, and what conditions are necessary for this to happen.*

***Keywords:*** *Holy See; algorethics; artificial intelligence; ethics; Pope Francis*

## Introducción

A diferencia de otros grandes campos de la investigación científica y del desarrollo tecnológico, la irrupción de los sistemas de inteligencia artificial que se ha producido en los últimos años ha puesto de manifiesto una cuestión ética. Esta urgencia no solo ha involucrado a los especialistas, como sucedió en el pasado reciente con el desarrollo de la investigación genética o en los años 70 con el auge de la investigación biológica, sino que se ha convertido en un tema ampliamente debatido en la comunidad científica, en el debate político, incluso en el sentir común, comparable quizás a lo que sucedió con el tema nuclear.

A esta proliferación del tema ético relacionado con los sistemas de inteligencia artificial ha contribuido, de manera significativa, una intensa actividad de la Santa Sede. El Papa Francisco, personalmente, y varios organismos vaticanos han abordado y debatido el tema en los últimos años, proponiendo una agenda de temas y una metodología de trabajo. Este artículo ofrece una presentación razonada y crítica de la contribución de la Santa Sede en la construcción de la actual reflexión ética occidental sobre la inteligencia artificial.

## Método

Los sujetos de la Santa Sede que participan de manera significativa y estructurada en el trabajo sobre la ética de la inteligencia artificial son, en primer lugar, el Pontífice, seguido de la Pontificia Academia para la Vida con la Fundación RenAIssance, el Dicasterio para la Cultura y la Educación y el Dicasterio para la Comunicación. Otras entidades vaticanas han abordado el tema de manera esporádica. El artículo examina las intervenciones e iniciativas de los sujetos relacionados con el tema de la ética de la inteligencia artificial, mostrando sus cuestiones fundamentales, los diferentes sujetos involucrados y la metodología de acción implementada.



## El magisterio del Papa Francisco

El magisterio del Papa Francisco sobre el tema de la inteligencia artificial consta, en primer lugar, de tres documentos principales, todos ellos de 2024: “Mensaje para la Jornada de la Paz”, “Mensaje para la Jornada de las Comunicaciones Sociales” y “Discurso ante el G7 en Borgo Ignazia”. Además de estos textos, especialmente estructurados, hay que reseñar una decena de discursos y mensajes, relacionados con iniciativas promovidas por otras realidades de la Santa Sede. En detalle y en orden cronológico:

- Discurso ante el Congreso “Child Dignity in the Digital World” (2019)
- Discurso ante la Pontificia Academia para la Vida sobre la algorética (2020)
- Discurso ante los líderes religiosos firmantes de la *Rome Call* (2023)

- Discurso ante la Pontificia Academia para la Vida sobre tecnología y vida (2023)
- Discurso ante la Sociedad Max Planck (2023)
- Discurso ante la Minerva Dialogues (2023)
- Discurso ante la Fundación Centesimus Annus (2024)
- Mensaje para el encuentro "AI Ethics for Peace" (2024)
- Discurso ante la Academia Pontificia de las Ciencias (2024)
- Mensaje ante la cumbre de París "para la acción sobre la inteligencia artificial" (2025).

Llama la atención, en primer lugar, el número y la importancia de las intervenciones del Papa Francisco, teniendo en cuenta que la tecnología y la inteligencia artificial no son temas tradicionalmente religiosos. De especial relevancia simbólica, más que de contenido, son el discurso ante el G7 (el Papa podría haber elegido hablar a los grandes de la Tierra sobre otros temas quizás más afines, como, por ejemplo, la libertad religiosa o la paz) y la acuñación de un neologismo, «algorética», que ha recibido un reconocimiento oficial y ha gozado de cierta difusión. El motivo por el que la Iglesia católica interviene en un tema tecnológico lo explica el Papa Francisco en el discurso ante la Pontificia Academia para la Vida de 2020:



> "La llamada 'inteligencia artificial' está en el corazón mismo del cambio de época que estamos atravesando. La innovación digital, efectivamente, alcanza a todos los aspectos de la vida, tanto personales como sociales. Afecta a la forma en que entendemos el mundo y a nosotros mismos. Está cada vez más presente en las actividades e incluso en las decisiones humanas, y está cambiando nuestra forma de pensar y actuar […] En efecto, la profundidad y la aceleración de las transformaciones de la era digital plantean problemas inesperados que imponen nuevas condiciones al ethos individual y colectivo" (Francisco, 2020).

Lejos de caer en juicios alarmistas o incluso apocalípticos, el texto reconoce la capacidad transformadora de estas tecnologías, definidas poco después como "don de Dios", que impone una reflexión amplia y renovada, con el fin de salvaguardar la vida humana y los vínculos sociales, gracias a un nuevo compromiso en el ámbito ético y educativo. El Pontífice lleva a cabo esta reflexión con un triple marco de referencia: en primer lugar, la Doctrina Social de la Iglesia (dignidad de la persona, justicia, solidaridad y subsidiariedad), luego el díptico de las dos encíclicas *Laudato Si'* (sobre el cuidado de la casa común) y *Fratelli tutti* (sobre la fraternidad universal), citadas en varias ocasiones en muchos discursos, y, por último, los derechos humanos que "en el encuentro entre diferentes visiones del mundo, constituyen un importante punto de convergencia para la búsqueda de un terreno común" (Francisco, 2024b).

En el discurso ante el G7, el Papa Francisco (2024b) argumenta la necesidad de una inspiración ética de la inteligencia artificial a partir de dos consideraciones fundamentales: la peculiaridad de este instrumento definido como *sui generis* por su complejidad y autonomía, la no neutralidad de cualquier innovación tecnológica y el consiguiente impacto en la vida y en la sociedad humanas.

> "No debemos olvidar que ninguna innovación es neutral. La tecnología nace con un propósito y, en su impacto en la sociedad humana, representa siempre una forma de orden en las relaciones sociales y una disposición de poder, que habilita a alguien a realizar determinadas acciones impidiéndoselo a otros. Esta dimensión de poder que es constitutiva de la tecnología incluye siempre, de una manera más o menos explícita, la visión del mundo de quien la ha realizado o desarrollado. Esto vale también para los programas de inteligencia artificial. Con el fin de que estos instrumentos sean para la construcción del bien y de un futuro mejor, deben estar siempre ordenados al bien de todo ser humano. Deben contener una inspiración ética. La decisión ética, de hecho, es aquella que tiene en cuenta no sólo los resultados de una acción, sino también los valores en juego y los deberes que se derivan de esos valores" (Francisco, 2024b).



El bien del ser humano evocado por el Pontífice tiene, ante todo, en la custodia y centralidad de la persona humana, que debe ser custodiada y no reducida o marginada, uno de los pilares de la reflexión ética sobre la inteligencia artificial, como se desprende de la repetición de esta intuición en muchos de sus discursos. "Nuestro último desafío es y seguirá siendo siempre el hombre", escribe el Papa Francisco (2025) a Macron, y un tiempo antes:

> "Desde esta perspectiva, creo que el desarrollo de la inteligencia artificial y del aprendizaje automático tiene el potencial de aportar una contribución beneficiosa al futuro de la humanidad, no podemos descartarlo. Sin embargo, estoy seguro de que este potencial sólo se hará realidad si existe una voluntad coherente por parte de quienes desarrollan las tecnologías para actuar de forma ética y responsable. Me anima el compromiso de tantas personas que trabajan en estos campos para garantizar que la tecnología se centre en el ser humano, se funde en bases éticas durante el diseño del proyecto y tenga por finalidad el bien […] El concepto de dignidad humana -este es el núcleo- exige que reconozcamos y respetemos el hecho de que el valor fundamental de una persona no puede medirse con un conjunto de datos […] No podemos permitir que los algoritmos limiten o condicionen el respeto a la dignidad humana, ni que excluyan la

> compasión, la misericordia, el perdón y, sobre todo, la apertura a la esperanza de cambio en el individuo" (Francisco, 2023d).

Además de la centralidad de la persona humana, el Papa Francisco (2020) plantea como segundo tema ético central el de las repercusiones sociales de la implementación de los sistemas de inteligencia artificial, mostrando en varios pasajes cómo este fenómeno está aumentando la injusticia y las desigualdades en el mundo en lugar de reducirlas, reduciendo los espacios de democracia y alimentando una cultura de la eficiencia que aumenta la brecha de los más débiles y frágiles.

Los diferentes discursos están plagados de ejemplos concretos de estas repercusiones sociales. Dependiendo de los contextos, el Pontífice recuerda los riesgos que se vislumbran en el mundo laboral, en la medicina, en los fenómenos migratorios, en la práctica de la justicia, en la educación, en la elaboración de perfiles de las personas, en el control y la calificación social, en la seguridad. El Pontífice dedica especial atención a los temas de la paz y de la comunicación en la era de la inteligencia artificial, que condensa en dos documentos específicos.

En el largo mensaje para el Día de la Paz 2024 (Francisco, 2023e), el Papa Francisco dedica un capítulo a los "temas candentes para la ética", abordando brevemente cuestiones como la fiabilidad de los sistemas, un sistema generalizado de manipulación y control social, un cambio en el mundo laboral que aumenta la injusticia social. En este contexto, el Pontífice afirma que:



> "La búsqueda de las tecnologías emergentes en el sector de los denominados 'sistemas de armas autónomos letales', incluido el uso bélico de la inteligencia artificial, es un gran motivo de preocupación ética. Los sistemas de armas autónomos[1] no podrán ser nunca sujetos moralmente responsables. La exclusiva capacidad humana de juicio moral y de decisión ética es más que un complejo conjunto de algoritmos, y dicha capacidad no puede reducirse a la programación de una máquina que, aun siendo 'inteligente', no deja de ser siempre una máquina. Por este motivo, es imperioso garantizar una supervisión humana adecuada, significativa y coherente de los sistemas de armas" (Francisco, 2023e).

En clave positiva, el Pontífice declara que:

[1] El tema de las armas autónomas también se retoma, con la misma preocupación, en el discurso ante el G7 (Francisco, 2024b) y en "Mensaje para el encuentro de Hiroshima" (Francisco, 2024d).

> "Realmente lo último que el mundo necesita es que las nuevas tecnologías contribuyan al injusto desarrollo del mercado y del comercio de las armas, promoviendo la locura de la guerra. […] Las aplicaciones técnicas más avanzadas no deben usarse para facilitar la resolución violenta de los conflictos, sino para pavimentar los caminos de la paz" (Francisco, 2023e).

El uso de la inteligencia artificial para construir un mundo más justo y pacífico –concluye el mensaje– requiere un esfuerzo educativo y comunicativo adecuado, así como un compromiso renovado de los actores políticos involucrados en las grandes decisiones globales.

"Mensaje para la Jornada de las Comunicaciones Sociales 2024" (Francisco, 2024a) explora con mayor precisión los desafíos que la inteligencia artificial plantea al gran mundo de la comunicación.[2] El mensaje comienza con un llamamiento a una visión sapiencial y espiritual de esta tecnología que, de alguna manera, invade el espacio propio del lenguaje humano, con el fin de preservar "una comunicación plenamente humana" que no se reduzca a un flujo de datos desencarnado. Aquí la lista de cuestiones sobre las que reflexionar es más detallada: explosión de noticias falsas, preponderancia de las redes sociales, riesgo de *echo chamber*, construcción de un pensamiento único, campañas de desinformación, peligros para los trabajadores del sector, reducción de los espacios de participación, aumento de las desigualdades sociales también debido a la gestión de las noticias. Ante estos peligros, el documento destaca:



> "El reto que tenemos ante nosotros es dar un salto cualitativo para estar a la altura de una sociedad compleja, multiétnica, pluralista, multirreligiosa y multicultural. Nos corresponde cuestionarnos sobre el desarrollo teórico y el uso práctico de estos nuevos instrumentos de comunicación y conocimiento" (Francisco, 2024a).

Esta dificultad, continúa el Pontífice, solo puede afrontarse manteniendo relaciones existenciales que impliquen los cuerpos, el arraigo en la realidad, compartir experiencias y no solo datos. En este sentido, parece decisiva la "alianza entre generaciones, entre quienes tienen memoria del pasado y quienes tienen visión de futuro". La algorética se convierte en la clave de síntesis para comprender el mensaje del Papa Francisco. El neologismo aparece por primera vez en los discursos del Papa en 2019 y se retoma varias veces en sus intervenciones. La algorética se define como una nueva frontera de la reflexión ética:

[2] Sobre el mismo tema, véase también el Documento del Dicasterio para la Comunicación (2023), dedicado específicamente al tema de las redes sociales, en el que la IA emerge como un factor de multiplicación de noticias peligrosamente problemático.

> "Su objetivo es asegurar una verificación competente y compartida de los procesos con los que se integran en nuestra era las relaciones entre los seres humanos y las máquinas [...] La 'algor-ética' podrá ser un puente para que los principios [de la Doctrina Social de la Iglesia] se inscriban concretamente en las tecnologías digitales, mediante un diálogo transdisciplinario eficaz" (Francisco, 2020).

Según Francisco (2023e), debe ofrecerse un desarrollo ético de los algoritmos, un marco de valores que pueda orientar de manera eficaz y concreta el desarrollo tecnológico, una "moderación ética" (Francisco, 2024b) que debe construirse ya en la fase de diseño (*by design*) con los ingenieros y las realidades industriales implicadas. Resulta especialmente interesante la mención del Pontífice sobre esta colaboración en un tema ético con las realidades científicas implicadas, que aparece en el discurso ante el G7:

> "En un contexto plural y global, en el que también se muestran las distintas sensibilidades y plurales jerarquías en las escalas de valores, parecería difícil encontrar una única jerarquía de valores. Pero en el análisis ético podemos recurrir además a otros tipos de instrumentos. Si nos cuesta definir un solo conjunto de valores globales, podemos encontrar principios compartidos con los cuales afrontar y disminuir eventuales dilemas y conflictos de la vida. Por esta razón ha nacido la *Rome Call*. En el término 'algorética' se condensa una serie de principios que se revelan como una plataforma global y plural capaz de encontrar el apoyo de las culturas, las religiones, las organizaciones internacionales y las grandes empresas protagonistas de este desarrollo" (Francisco, 2024b).



**La Pontificia Academia para la Vida y la Fundación RenAIssance**

La contribución ofrecida por la Pontificia Academia para la Vida[3] se inscribe en un proceso de reflexión científica más amplio, iniciado en 2016 con la reforma de la entidad deseada por el Papa Francisco. La academia destacó desde el principio la especial urgencia de la relación entre las nuevas tecnologías y la vida humana, iniciando un camino de investigación que la llevó a analizar, a través de una serie de congresos y publicaciones, el influjo de las tecnologías biomédicas en los momentos del nacimiento y la muerte, la redefinición del concepto de humano en relación con el avance de la robótica, el influjo poderoso de las llamadas tecnologías emergentes y convergentes.[4]

[3] Los objetivos, la estructura y las actividades de la Academia pueden verse en: www.academyforlife.va.
[4] Entre los frutos más prestigiosos de este trabajo se encuentra el artículo publicado en *Nature Machine Intelligence*: Sinibaldi *et al.*, 2020.

En este marco se inserta el congreso de 2020 (con las actas publicadas), *The Good Algorithm? Artificial Intelligence: Ethics, Law, Health* (Paglia, 2021b), y el lanzamiento asociado de la *Rome Call for AI Ethics*.

La sesión del congreso dedicada a la ética en el campo de la IA sirvió para sentar las bases del problema, que en aquel momento era objeto de estudio solo por parte de especialistas. A través de cuatro ponencias se puso de manifiesto la necesidad de abordar el tema en el marco más amplio de la transformación tecnológica y digital, la dimensión social y multicultural en la que debía situarse esta reflexión y las repercusiones educativas conexas. El acto de clausura del congreso fue la ceremonia de firma de la *Rome Call for AI Ethics*, inaugurada por el mensaje del Papa (Francisco, 2020), en la que participaron como firmantes: el director general de la FAO, Qu Dongyu; el presidente de Microsoft, Brad Smith; el vicepresidente de IBM, John Kelly III; la Ministra de Transición Tecnológica del Gobierno italiano, Paola Pisano; y el presidente de la Academia, monseñor Vincenzo Paglia. También asistió, sin firmar, el presidente del Parlamento Europeo, David Sassoli.

La *Rome Call*[5] es un documento breve, compuesto por un marco antropológico de fondo articulado en tres breves capítulos (ética, educación, derecho) y seis principios:

- *Transparencia:* en general, los sistemas de inteligencia artificial deben ser explicables;
- *Inclusión:* las necesidades de todos los seres humanos deben tenerse en cuenta para que todos puedan beneficiarse y para que se ofrezcan a todas las personas las mejores condiciones posibles para expresarse y desarrollarse;
- *Responsabilidad:* quienes diseñan y ponen en práctica estas tecnologías deben proceder con responsabilidad y transparencia;
- *Imparcialidad:* evitar crear o actuar según prejuicios, salvaguardando así la equidad y la dignidad humana;
- *Fiabilidad:* los sistemas de inteligencia artificial deben ser capaces de funcionar de manera fiable;
- *Seguridad y privacidad:* los sistemas de inteligencia artificial deben funcionar de forma segura y respetar la privacidad de los usuarios.



El texto, fruto del trabajo de una comisión de expertos de diversas disciplinas y culturas dirigida por el profesor Paolo Benanti, no es deliberadamente un texto vaticano y, por lo tanto, no forma parte en modo alguno del corpus del magisterio católico. Más bien es un texto propuesto por un organismo de la Iglesia católica a todos aquellos interesados en

[5] Para conocer el texto, todos los firmantes y las iniciativas relacionadas con la difusión del documento, se puede consultar la página oficial: www.romecall.org. El sitio está gestionado por la Fundación vaticana RenAIssance, creada por el Papa Francisco específicamente para custodiar, profundizar y difundir la *Rome Call* y, en general, un enfoque ético de la tecnología.

construir juntos una visión ética de la inteligencia artificial y dispuestos a hacer suya la visión del documento y suscribirlo. El único texto fundamental citado e inspirador es la Declaración Universal de los Derechos Humanos. Esta ubicación responde al método y al propósito del documento. De hecho, adopta el método indicado por el Papa Francisco de dialogar sobre estos temas con todos los sujetos involucrados y está formulado de manera que todos los sujetos que lo deseen y se comprometan en este sentido puedan suscribirlo, sin tener que declarar necesariamente una opción religiosa.

Desde el punto de vista de la difusión del documento, esta formulación ha sido especialmente eficaz. Hasta la fecha, lo han firmado varias otras industrias (entre ellas CISCO en 2024, y Qualcomm y Salesforce en 2025), decenas de universidades de todo el mundo y los líderes de todas las principales religiones del mundo.[6] El Papa Francisco ha definido la *Rome Call* como "una plataforma global y plural capaz de encontrar el apoyo de las culturas, las religiones, las organizaciones internacionales y las grandes empresas protagonistas de este desarrollo" (Francisco, 2024b). El *AI Index Report 2021* publicado por el Institute for Human-Centered AI de la Universidad de Stanford lo ha definido como uno de los cinco documentos más importantes sobre el tema en 2020. Amandeep Singh Gill, enviado para la tecnología del secretario general de las Naciones Unidas, declaró en su discurso en el evento de Hiroshima para la firma de la *Rome Call* por parte de los líderes religiosos del mundo: "*The Rome Call for AI Ethics embodies the spirit needed for global AI governance*".



En resumen, se puede afirmar que la Pontificia Academia ha desempeñado cuatro funciones en el campo del desarrollo de la ética de la IA. En primer lugar, ha llamado la atención, antes que muchos otros, sobre un tema que luego se ha convertido en central en el debate académico y público. En segundo lugar, situó esta cuestión en una reflexión antropológica y social más amplia, sin limitarse a ofrecer, como muchos otros que le siguieron, una mera lista de principios reguladores. Benanti sitúa la algorética en el marco de una condición que él define como tecno-humana:

> "*The way humans have always understood and lived their existence: an interaction with the environment mediated through tools, technological artifacts. It seems clear, then, that it is not possible to separate the history of human beings and civilization from the history of the tools they made*" (Benanti, 2023).

En tercer lugar, ha intentado aplicar un método transdisciplinario en el que el resultado final es fruto del trabajo conjunto de expertos pertenecientes a ámbitos científicos muy diferentes. De hecho, ha conseguido dar forma a lo que en varias ocasiones

[6] En 2023, líderes religiosos judíos y musulmanes firmaron juntos la *Rome Call* en el Vaticano. En 2024, en Hiroshima, líderes budistas, sijs, confucianos, hindúes, bahaíes y zoroastrianos firmaron el documento. Para un análisis detallado de la participación de las religiones, véase: Ciucci (2025).

había sido solicitado por el mundo tecnológico, que denunciaba la problemática falta, en sus entornos, de un punto de vista humanista. Por último, la *Rome Call* ha sido un poderoso medio de difusión del tema y su visión en mundos y contextos muy diferentes. Paglia escribe sucintamente:

> "*Possiamo cogliere, anche se fra molte contraddizioni e tensioni, la graduale maturazione di un comune impegno per un autentico sviluppo, che intende il progresso in termini non solo economici o quantitativi. Il pur indispensabile contributo del sapere tecnico-scientifico viene infatti articolato con altre prospettive, incluse quelle della teologia e dell'etica, in modo da orientarci, con la collaborazione di tutte le persone di buona volontà, verso uno sviluppo umano effettivamente integrale*" (Paglia, 2021a).

**El Dicasterio para la Cultura y la Educación**

El otro gran actor vaticano que se ha movido en el tema de la ética de la inteligencia artificial es sin duda el Dicasterio para la Cultura y la Educación. Este compromiso se ha desarrollado en cuatro direcciones fundamentales: la participación en congresos científicos y debates de algunas figuras de alto nivel del dicasterio, la promoción de los *Minerva Dialogues*,[7] la constitución de un grupo de trabajo dedicado al tema que produjo en 2024 un volumen sobre el tema (Gaudet *et al.*, 2024) y la publicación en 2025, en colaboración con el Dicasterio para la Doctrina de la Fe, de la Nota *Antiqua et Nova*.



Si los dos primeros elementos muestran la adquisición efectiva del método que consiste en escuchar y dialogar, como pidió el Papa Francisco en su discurso a los participantes en los *Minerva Dialogues* de 2023, los dos documentos proponen un análisis del fenómeno de la inteligencia artificial y esbozan algunas perspectivas de carácter antropológico y ético. El volumen *Encountering Artificial Intelligence* dedica la segunda parte al tema de la ética:

> "*In this part of the volume, we respond to Pope Francis's call with an outline of an AI ethics that can assist individuals, institutions, and civil society in engaging personal, social, and political concerns related to AI in light of the Catholic theological and moral tradition*" (Gaudet *et al.*, 2024, p. 147).

[7] Los *Minerva Dialogues* son encuentros anuales, iniciados en 2016 por iniciativa del fraile dominico Éric Salobir. Como explica el secretario del Dicasterio, Paul Tighe, en un artículo publicado en el *Osservatore Romano* (30 de marzo de 2023), constituyen una oportunidad para que expertos del mundo de la tecnología –entre los que se encuentran científicos, ingenieros, directivos de empresas, juristas y filósofos– se reúnan con representantes de la iglesia, como oficiales de la curia, teólogos y moralistas. Nacidos para promover una mayor conciencia y reflexión sobre el impacto social y cultural de las tecnologías digitales, según Tighe, ofrecen la oportunidad de "*indagare insieme per capirci meglio, per assicurarci che la tecnologia sia applicata per il bene dell'umanità, parlando anche con schiettezza in certi momenti*".

En el capítulo seis, gracias a un análisis general del magisterio del Papa Francisco (definido como positivo pero no ingenuo), el volumen radica la cuestión de la ética de la inteligencia artificial en el ámbito más amplio de la doctrina social de la Iglesia, mostrando cómo las categorías fundamentales de esta reflexión pueden abordar positivamente este nuevo fenómeno: bien común, solidaridad, subsidiariedad, libertad, educación y familia son las palabras que permiten identificar y asegurar un impacto positivo de la inteligencia artificial en la vida personal y social de las personas.

> "*The goal of chapter 7 is to identify how AI systems can make the realization of the ideal even more difficult. To serve this purpose, the chapter articulates the vision and the challenges at the same time. We will begin with how AI is shaping more intimate spheres of encounter, such as the family, before expanding to look at AI's effects on our broader social systems, like government and the economy*" (Gaudet *et al.*, 2024, p. 161).

A través de las categorías identificadas en el capítulo anterior, el volumen analiza los beneficios y riesgos –introducidos por los sistemas de inteligencia artificial– sobre la familia (cuyas relaciones están moldeadas por las nuevas tecnologías), sobre la educación y el paradigma tecnocrático, sobre la salud y la medicina, sobre la protección de los más débiles (arriesgadamente expuestos a la eficiencia de los algoritmos), sobre el mundo laboral con la pérdida y la creación de figuras profesionales, sobre el ámbito militar con especial atención a las armas autónomas, sobre los procesos culturales y sobre el impacto medioambiental de estos sistemas.



Por último, el capítulo ocho ofrece una visión sapiencial en la que es posible habitar un tiempo digital, destacando toda una serie de ejercicios de humanidad y atenciones que deben preservarse en la vida cotidiana (que no deben digitalizarse y autonomizarse indiscriminadamente), en el cuidado de las personas y situaciones más vulnerables (que deben encontrar en las nuevas tecnologías un beneficio y no un enésimo factor de exclusión), en el debate público y jurídico (preservando la libertad de las personas y valores como la equidad, la privacidad, la seguridad y la responsabilidad), en la economía (llamada a custodiar un rostro humano) y en las relaciones interpersonales (que no pueden reducirse a formas exclusivamente digitales). En cada uno de estos temas se presentan también puntos de trabajo para quienes diseñan y crean sistemas de inteligencia artificial.

## La nota *Antiqua et Nova*

El 14 de enero de 2025, el Dicasterio para la Cultura y la Educación, junto con el Dicasterio para la Doctrina de la Fe, publicó *Antiqua et Nova - Nota sobre la relación entre la inteligencia artificial y la inteligencia humana*. El extenso documento ofrece

una reflexión articulada sobre el tema de la inteligencia artificial en diálogo con la tradición cristiana y recoge todo el trabajo realizado por la Santa Sede sobre el tema.

Si en las tres primeras partes el documento enmarca el tema de la inteligencia artificial en comparación con la inteligencia humana, la cuarta parte está dedicada específicamente a la cuestión ética. Gracias a este marco, al final, la larga quinta parte retoma varias cuestiones específicas que destacan, retomando el esquema y el índice del volumen *Encountering Artificial Intelligence*, las posibilidades y los riesgos del uso de la inteligencia artificial en los diferentes campos de la experiencia humana. Los primeros números del cuarto volumen definen el marco en el que se debe plantear la cuestión ética en este tema. La inteligencia artificial es una empresa humana y, como toda actividad humana, pide ser comprendida dentro del designio de Dios: el intelecto es un don suyo para ser usado para el bien (Dicasterio Doctrina, 2025, pp. 36-38).

El documento destaca dos características fundamentales. En primer lugar, el resultado de la comparación entre la inteligencia humana y la artificial muestra que la ética de la inteligencia artificial es, en realidad, siempre la ética de las personas:

> "La dimensión ética es primordial, ya que son las personas las que diseñan los sistemas y determinan para qué se utilizan. Entre una máquina y un ser humano, sólo este último es verdaderamente un agente moral, es decir, un sujeto moralmente responsable que ejerce su libertad en sus decisiones y acepta las consecuencias de las mismas; sólo el ser humano está en relación con la verdad y el bien, guiado por la conciencia moral que le llama a "amar y practicar el bien y que debe evitar el mal", certificando "la autoridad de la verdad con referencia al Bien supremo por el cual la persona humana se siente atraída"; sólo el ser humano puede ser lo suficientemente consciente de sí mismo como para escuchar y seguir la voz de la conciencia, discerniendo con prudencia y buscando el bien posible en cada situación. En realidad, esto también pertenece al ejercicio de la inteligencia por parte de la persona" (Dicasterio Doctrina, 2025, p. 39).



En segundo lugar, la custodia de la dignidad de la persona humana y del bien común (los fines de la acción ética también en este campo, como se señala en los números 42 y 48) deben ser perseguidos tanto por quienes utilizan estas tecnologías (custodiando sus fines, como se indica en el número 40) como por quienes diseñan y construyen estos sistemas (los medios no son moralmente indiferentes, como se indica en el número 41). En esencia, una ética *by design*, como ha señalado en varias ocasiones el Papa Francisco y se destaca en la *Rome Call*.

La segunda mitad de la parte dedicada a la ética subraya, por último, una cuestión que requiere especial atención:

> "Dado que una causalidad moral en sentido pleno sólo pertenece a los agentes personales, no a los artificiales, es de suma importancia poder identificar y definir quién es responsable de los procesos de IA, en particular de aquellos que incluyen posibilidades de aprendizaje, corrección y reprogramación. Si bien, por un lado, los métodos empíricos (*bottom-up*) y las redes neuronales muy profundas permiten a la IA resolver problemas complejos, por otro lado, dificultan la comprensión de los procesos que condujeron a tales soluciones. Esto complica la determinación de responsabilidades ya que, si una aplicación de IA produjera resultados no deseados, sería difícil determinar a qué persona atribuirlos" (Dicasterio Doctrina, 2025, 44).

Aquí, en el documento, aparece el tema de la *accountability* y de la necesidad de instrumentos jurídicos reguladores adecuados, capaces de salvaguardar y reconocer al mismo tiempo las responsabilidades humanas, incluso en procesos tecnológicos y estructuras sociales extremadamente complejos.

## Conclusiones



El largo análisis dedicado a las intervenciones e iniciativas promovidas por la Santa Sede en los últimos cinco años sobre el tema de la ética de la inteligencia artificial ofrece algunas conclusiones.

La primera y fundamental, porque ha hecho posibles las demás, es de orden metodológico. La dirección fue claramente dada por el Papa Francisco, quien ya en 2020 afirmó:

> "En efecto, como creyentes, no tenemos nociones preestablecidas con las que responder a las preguntas sin precedentes que la historia hoy nos plantea. Nuestra tarea es, más bien, caminar junto con los demás, escuchando atentamente y poniendo en contacto la experiencia y la reflexión. Debemos dejarnos interpelar como creyentes, para que la Palabra y la Tradición de la fe nos ayuden a interpretar los fenómenos de nuestro mundo, identificando caminos de humanización, y por tanto de evangelización amorosa, para recorrerlos juntos. Así podremos dialogar provechosamente con todos aquellos que buscan el desarrollo humano, manteniendo a la persona en todas sus dimensiones, incluidas las espirituales, en el centro del conocimiento y las prácticas sociales. Nos enfrentamos a una tarea que involucra a la familia humana en su totalidad" (Francisco, 2020).

Todas las intervenciones del Vaticano han seguido esta modalidad de diálogo y apertura con todos los sujetos involucrados. En este sentido, por ejemplo, la Iglesia Católica ha optado por no designar la *Rome Call* como un documento vaticano.

Al seguir este método, la Santa Sede ha ocupado el espacio público no para defender una visión del mundo, algunas de sus prerrogativas o, peor aún, algunas posiciones privilegiadas, sino para poner a disposición del debate público una visión articulada del hombre y de la sociedad. En cierto modo, sucedió exactamente lo contrario de lo que ocurrió en el episodio más llamativo de enfrentamiento entre la iglesia y los científicos de la época moderna: el caso Galileo. En aquella ocasión, los cardenales que trataron el asunto se opusieron a Galileo no tanto por razones científicas, sino más bien porque intuían que la aceptación de lo propuesto por el científico toscano socavaría gravemente la visión teológica de la época, que situaba en el centro del universo el acontecimiento cristiano y la consiguiente función de la Iglesia. En el caso de Galileo, la Iglesia se atrincheró en un intento apologético; hoy, ante la explosión de la inteligencia artificial, el Santo Padre ha salido al ágora pública para contribuir a la reflexión común, sin "nociones preconcebidas con las que responder a las preguntas sin precedentes que la historia hoy nos plantea".

La ampliación de los sujetos implicados ha supuesto asumir algún riesgo: ha surgido cierta perplejidad en torno a la participación de algunas de las mayores empresas mundiales del sector tecnológico y de sus líderes. La decisión de trabajar con ellos, tomada después de largos períodos de reflexión y discernimiento, se deriva de la percepción de que una reflexión ética que no involucrara a los actores y autores de esta revolución digital privaría al proyecto de consistencia y aplicabilidad reales. Las divergencias, incluso profundas, sobre aspectos de carácter principalmente social no impidieron encontrar convergencias y esperanzas comunes sobre las que trabajar juntos, por el bien de todos.



La segunda conclusión se refiere al juicio general que la Santa Sede da al fenómeno de la inteligencia artificial. Como se desprende del análisis realizado en las páginas anteriores, en ningún texto aparece un tono apocalíptico y de grave juicio condenatorio. También son muy escasas las referencias al poshumanismo y al transhumanismo, tradiciones de pensamiento que a menudo se asocian a juicios extremadamente negativos sobre la transición digital en curso. Ciertamente, la Santa Sede denuncia el paradigma tecnocrático "que promete un progreso incontrolado e ilimitado se impondrá y quizás incluso eliminará otros factores de desarrollo con enormes peligros para toda la humanidad" (Francisco, 2019). Pero el juicio sobre la inteligencia artificial es positivo y, al mismo tiempo, no ingenuo; reconoce esta tecnología como uno de los mejores frutos del hombre y, al mismo tiempo, destaca sus riesgos y oportunidades.

La algorética no pretende simplemente enumerar una serie de principios reglamentarios destinados a reducir los riesgos del impacto de esta tecnología; más bien, inserta la inteligencia artificial en un marco prospectivo positivo. La pregunta que surge es: ¿cómo podemos, juntos, construir y utilizar estos sistemas para construir

el bien de las personas, de la sociedad, del planeta?[8] La idea de una tecnología no neutral y la atención a sus repercusiones sociales, y no solo antropológicas, se derivan de este enfoque y caracterizan de manera peculiar la reflexión ética propuesta por la Santa Sede sobre el tema. La pregunta que permanece en los corazones y mentes de muchos es la relativa a la eficacia real que todo este camino puede tener, en un contexto global caracterizado por inmensos intereses económicos y políticos. La discusión de los últimos años ha mostrado un primer resultado: ciertamente, la Santa Sede ha contribuido de manera significativa a la construcción y promoción del tema de la ética de la inteligencia artificial, y ha sido esencial para su posicionamiento entre las grandes cuestiones de la discusión pública mundial. No es poca cosa y no ocurre con otros sectores, igualmente decisivos, de la investigación científica y tecnológica. En un futuro próximo, será decisivo ampliar la conciencia de la urgencia de la cuestión entre los diferentes sujetos implicados y la capacidad de los políticos y constructores de encontrar reglas y soluciones capaces de apoyar al mismo tiempo la justicia y el desarrollo, la protección y la innovación.

Por último, para la Iglesia católica otra prueba de fuego será la gestión de la transformación digital en la práctica pastoral de las distintas iglesias locales. Reconocer el don de Dios de la inteligencia artificial y no pensar que esto también afecta a las condiciones concretas de la práctica eclesial podría ser el primer fracaso.



**Bibliografía**

Benanti, P. (2020). Algor-Ethics: Artificial Intelligence and Ethical Reflection. Revue d'éthique et de théologie morale, 307(3), 93-110.

Benanti, P. (2023). The urgency of an algorethics. Discovering Artificial Intelligence, 3(11). DOI: https://doi.org/10.1007/s44163-023-00056-6.

Ciucci, A. (2024). Le religioni e l'etica dell'intelligenza artificiale. Il caso della Rome Call for AI Ethics. La Società, 5-6, 100-109.

Dicasterio para la Comunicación (2023). Hacia una plena presencia - Reflexión pastoral sobre la interacción en las Redes Sociales. Recuperado de: https://www.vatican.va/roman_curia/dpc/documents/20230528_dpc-verso-piena-presenza_es.html.

Dicasterio para la Doctrina de la Fe y Dicasterio para la Cultura y la Educación (2025). Antiqua et Nova. Nota sobre la relación entre la inteligencia artificial y la inteligencia

[8] En este sentido, se puede pensar que el marco ético propuesto por la Santa Sede puede atribuirse, al menos en cierto modo, al sociotecnicismo pragmático defendido por Watson, Mokander y Floridi (2024), ya que no se corresponde ni con el optimismo ingenuo del sociotecnicismo dogmático ni con el pesimismo riguroso del escepticismo, considerados insuficientes por los autores.

humana. Recuperado de: https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_sp.html.

Gaudet, M., Herzfeld, N., Scherz, P. & Wales, J. (2024). Encountering Artificial Intelligence: Ethical and Anthropological Explorations. Eugene: Pickwick Publications.

Francisco (2019). Discurso a los participantes en el congreso "Promoting Digital Child Dignity". Recuperado de: https://www.vatican.va/content/francesco/es/speeches/2019/november/documents/papa-francesco_20191114_convegno-child%20dignity.html.

Francisco (2020). Discurso a los participantes en la plenaria de la Pontificia Academia para la Vida. Recuperado de: https://www.vatican.va/content/francesco/es/speeches/2020/february/documents/papa-francesco_20200228_accademia-perlavita.html.

Francisco (2023a). Discurso a los participantes en el encuentro "Rome Call" organizado por la Fundación RenAIssance. Recuperado de: https://www.vatican.va/content/francesco/es/speeches/2023/january/documents/20230110-incontro-romecall.html.

Francisco (2023b). Discurso a los Miembros de la Pontificia Academia para la Vida, recuperado de https://www.vatican.va/content/francesco/es/speeches/2023/february/documents/20230220-pav.html

Francisco (2023c). Discorso a una delegazione della Società Max Planck. Recuperado de: https://www.vatican.va/content/francesco/it/speeches/2023/february/documents/20230223-delegazione-mplanck.html.



Francisco (2023d). Discurso a los participantes en los "Minerva Dialogues" organizado por el Dicasterio para la Cultura y la Educación. Recuperado de: https://www.vatican.va/content/francesco/es/speeches/2023/march/documents/20230327-minerva-dialogues.html.

Francisco (2023e). Mensaje para la celebración de la 57 Jornada mundial de la Paz: Inteligencia artificial y paz. Recuperado de: https://www.vatican.va/content/francesco/es/messages/peace/documents/20231208-messaggio-57giornatamondiale-pace2024.html.

Francisco (2024a). Mensaje para la 58 Jornada mundial de las Comunicaciones Sociales: Inteligencia artificial y sabiduría del corazón para una comunicación plenamente humana. Recuperado de: https://www.vatican.va/content/francesco/es/messages/communications/documents/20240124-messaggio-comunicazioni-sociali.html.

Francisco (2024b). Discurso en la Sesión del G7 sobre inteligencia artificial. Recuperado de: https://www.vatican.va/content/francesco/es/speeches/2024/june/documents/20240614-g7-intelligenza-artificiale.html.

Francisco (2024c). Discorso ai partecipanti al convegno internazionale promosso dalla Fondazione Centesimus Annus Pro Pontifice. Recuperado de: https://www.vatican.va/content/francesco/it/speeches/2024/june/documents/20240622-centesimus-annus-propontifice.html.

Francisco (2024d). Mensaje para la reunión “AI Ethics for Peace”. Recuperado de https://www.vatican.va/content/francesco/es/messages/pont-messages/2024/documents/20240710-messaggio-ai-ethics-forpeace.html.

Francisco (2024e). Discurso a los participantes en la plenaria de la Academia de les Ciencias, recuperado de: https://www.vatican.va/content/francesco/es/speeches/2024/september/documents/20240923-plenaria-accademia-scienze.html.

Francisco (2025). Mensaje al Presidente de la República francesa con ocasión del “Sommet pour l’action sur l’intelligence artificielle”. Recuperado de: https://www.vatican.va/content/francesco/es/messages/pont-messages/2025/documents/20250207-messaggio-summit-parigi-ia.html.

Paglia, V. (2021a). Tecnologie digitali ed etica. Rivista Di Scienze dell’Educazione, 59(1), 68–80.

Paglia V. Pegoraro R. (2021b). The Good Algorithm? Artificial Intelligence: Ethics, Law, Health. Recuperado de: https://www.academyforlife.va/content/dam/pav/documenti%20pdf/2020/Assemblea/Abstract/Abstract%20book_GOLD%20(1).pdf.

Sinibaldi, E. *et al.* (2020). Contributions from the Catholic Church to ethical reflections in the digital era. Nature Machine Intelligence, 2, 242-244.

Watson, D. S., Mökander, J. & Floridi, L. (2024). Competing narratives in AI ethics: a defense of sociotechnical pragmatism. AI & Society.

# ARTÍCULOS SELECCIONADOS

Los artículos publicados en esta sección fueron seleccionados en el marco de una convocatoria abierta llevada adelante entre 2024 y 2025 por la Organización de Estados Iberoamericanos (OEI) para la Educación, la Ciencia y la Cultura, y la Pontificia Comisión para América Latina (PCAL). La selección se realizó sobre más de 60 artículos enviados desde Argentina, España, Chile, México, Perú, Colombia, Costa Rica, Ecuador, Brasil, Uruguay, Paraguay, Cuba y Venezuela. El trabajo de selección estuvo a cargo de Ismael Gómez, director de estrategia digital global de la OEI, y Diego Álvarez Newman, investigador del CONICET en la Universidad Nacional de José Clemente Paz (UNPAZ), Argentina, y miembro del Grupo de Estudios sobre IA del Consejo Episcopal Latinoamericano (CELAM). El comité de lectura estuvo integrado, además, por Karina Pedace (Universidad Nacional de La Matanza, Argentina), Betty Pérez Cedeño (Universidad de La Laguna, España), Luis Jiménez Rodríguez (Pontificia Universidad Católica de Puerto Rico) y Tobías Schleider (Universidad Nacional del Sur, Argentina).

# IA y política social. Discusiones éticas del SINIRUBE en Costa Rica

# IA e política social. Discussões éticas do SINIRUBE na Costa Rica

# *AI and Welfare Policy. SINIRUBE's Ethical Discussions in Costa Rica*

**Roberto Cascante Vindas** *



La política social costarricense dispone actualmente del Sistema Nacional de Información y Registro Único de Beneficiarios del Estado (SINIRUBE), que incorpora el uso de Inteligencia Artificial (IA) para implementar una única forma de medición de pobreza de carácter predictivo. A pesar del amplio desarrollo normativo y operativo efectuado, la lectura ética del SINIRUBE, según lo indicado en la *Recomendación sobre la ética de la inteligencia artificial* de la UNESCO y el documento *Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina y el Caribe*, de la OCDE y la CAF, así como los aportes deontológicos de las personas profesionales en trabajo social que participan en su ejecución, permite comprender los dilemas éticos asociados con la opacidad de su algoritmo, la desinformación respecto a su funcionamiento y la deshumanización del proceso a los que hacen frente los agentes morales que participan del sistema, así como poner en tela de juicio el cumplimiento de los derechos humanos, sobre todo para el caso de las poblaciones excluidas de los servicios asistenciales.

**Palabras claves**: derechos humanos; bienestar social; pobreza; deontología

* Bachiller y licenciado en trabajo social, Universidad de Costa Rica (UCR). Magíster en gerencia de proyectos de desarrollo, Instituto Centroamericano de Administración Pública, y doctorando en ciencias sociales sobre América Central (UCR). Docente e investigador de la carrera de trabajo social, Sede de Occidente (UCR), y de la Cátedra en Trabajo Social, Universidad Estatal a Distancia. Correo electrónico: roberto.cascantevindas@ucr.ac.cr. ORCID: https://orcid.org/0000-0001-7587-2821.

A política social da Costa Rica conta atualmente com o Sistema Nacional de Información y Registro Único de Beneficiarios (SINIRUBE), sistema que incorporou o uso da Inteligência Artificial (IA) para implementar uma forma única de medição preditiva da pobreza. Apesar do amplo desenvolvimento regulatório e operacional realizado, a leitura ética do SINIRUBE conforme indicada na *Recomendação sobre Ética da Inteligência Artificial* da Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO), o documento *Uso estratégico e responsável da inteligência artificial na o setor público da América Latina e do Caribe* da Organização para a Cooperação e o Desenvolvimento Econômico (OCDE) e a Corporação Andina de Desenvolvimento (CAF), bem como as contribuições deontológicas das pessoas Os profissionais de Serviço Social que participam na sua execução permitem-nos compreender dilemas éticos associados à opacidade e caixa negra do seu algoritmo, à desinformação quanto ao seu funcionamento e à desumanização do processo enfrentado pelos agentes morais participantes nas diferentes fases do sistema. Assim como o cumprimento dos direitos humanos é questionado, especialmente para as populações excluídas dos serviços de assistência social.

**Palavras-chave**: direitos humanos; assistência social; pobreza; deontología

*Costa Rican social policy currently operates through the National Information System and Single Registry of State Beneficiaries (SINIRUBE, due to its initials in Spanish), which incorporates the use of Artificial Intelligence (AI) to implement a unique form of predictive poverty measurement. Despite the extensive normative and operational development carried out, the ethical reading of SINIRUBE, as indicated in UNESCO's* Recommendation on the Ethics of Artificial Intelligence *and OECD and CAF's* Strategic and Responsible Use of Artificial Intelligence in the Public Sector in Latin America and the Caribbean*, as well as the deontological contributions of professionals in social work involved in its implementation, allows us to understand the ethical dilemmas associated with the opacity of its algorithm, the misinformation regarding its operation and the dehumanization of the process faced by moral agents involved in the system, as well as to question the fulfillment of human rights, especially in the case of populations excluded from assistance services.*

***Keywords**: human rights; social policy; poverty; deontology*

## Introducción

El estudio de las ciencias computacionales ha mostrado interés desde los años 60 del siglo XX por automatizar la conducta inteligente de máquinas y alcanzar resultados similares a los obtenidos por el razonamiento humano; más recientemente las herramientas que hacen uso de la inteligencia artificial (IA) han llegado a nutrirse de otros campos del conocimiento tales como la filosofía, la lingüística, las matemáticas y la psicología (Ponce & Torres, 2014). A la par, la conceptualización de la IA también ha cambiado, transitando desde posicionamientos que la catalogan como una ciencia propia, una rama o un estudio. Más recientemente, en su *Recomendación del Consejo sobre Inteligencia Artificial*, la Organización para la Cooperación y el Desarrollo Económico (OCDE) define los sistemas de IA como:

> "[...] un sistema basado en máquinas que, para objetivos explícitos o implícitos, infiere, a partir de la información que recibe (input), cómo generar resultados como predicciones, contenidos, recomendaciones o decisiones que pueden influir en entornos físicos o virtuales. Los distintos sistemas de IA varían en sus niveles de autonomía y adaptabilidad tras su despliegue" (OCDE, 2024, p. 7).



Por su parte, la UNESCO (2021) dentro de su *Recomendación sobre la ética de la inteligencia artificial*: "[…] considera los sistemas de IA sistemas capaces de procesar datos e información de una manera que se asemeja a un comportamiento inteligente, y abarca generalmente aspectos de razonamiento, aprendizaje, percepción, predicción, planificación o control" (p. 10). Tal ha sido el auge de la IA que actualmente se encuentra presente en la identificación de patrones de datos, toma de decisiones y procesos de selección de personal en empresas (Goñi, 2019), así como en labores cotidianas realizadas por el ser humano para autocompletar palabras o frases en teléfonos y correo electrónico, sistemas de traducción automatizada, programas para detectar errores ortográficos, e incluso en el uso de *chatbots* en procesos de investigación, entre otros ejemplos (Gutiérrez, 2023).

A pesar del uso de la IA en el sector privado y en la cotidianidad de las personas, su implementación en el diseño, la ejecución y la operacionalización de políticas estatales es un ámbito aún más reciente. Esto se da sobre todo en los lineamientos desplegados por la OCDE y la Corporación Andina de Fomento (CAF) (2022) en su documento *Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina y el Caribe*, y en los modelos de regulación ofrecidos por la UNESCO y la Unión Europea, específicamente a partir de su Ley de Inteligencia Artificial (Megías, 2022; Abdala *et al.*, 2019).

En el caso de América Latina, tanto la OCDE y la CAF (2022) como Jay *et al.* (2024) reconocen un avance en políticas públicas que bajo la connotación de modernización

hacen uso de la IA en la agilización de procesos considerados burocráticos.[1] En Costa Rica esto se comprueba en la adhesión del país a los principios de la OCDE respecto a IA (OCDE & CAF, 2022). Además, en 2023 el Ministerio de Ciencia, Innovación, Tecnología y Telecomunicaciones (MICITT) firmó una carta de compromiso con la UNESCO para contar con una estrategia nacional en materia de IA, de acuerdo con lo indicado en la *Recomendación sobre la ética de la inteligencia artificial* (Álvarez, 2023). Esto se materializó en 2024 con la *Estrategia Nacional de Inteligencia Artificial de Costa Rica*. Los ejes estratégicos de dicha estrategia para el periodo 2024-2027 se sintetizan en:

- Eje 1: IA ética, segura y responsable
- Eje 2: Articulación territorial y desarrollo económico
- Eje 3: Promoción de la I+D+I
- Eje 4: Gobierno inteligente
- Eje 5: Capacitación y formación de talento
- Eje 6: Infraestructura digital y tecnologías habilitantes
- Eje 7: Liderazgo internacional (MICITT, 2024)

Por otra parte, Costa Rica cuenta en corriente legislativa con tres proyectos de ley que están en discusión dentro de comisiones: la Ley de Regulación de la Inteligencia Artificial en Costa Rica (Castro *et al.*, 2023), la Ley para la promoción responsable de la Inteligencia Artificial en Costa Rica (Izquierdo, 2023) y la Ley para la Implementación de Sistemas de Inteligencia Artificial (IA) (Obando, 2024). En los tres expedientes se hace mención a la relevancia de articular la IA con los derechos humanos y la ética, entre otros aspectos.



Tal como se evidencia, se han planteado marcos normativos a nivel internacional y nacional para regular la IA, los cuales hacen alusión a discusiones éticas asociadas con los beneficios y riesgos para el cumplimiento de los derechos humanos (Megías, 2022). No obstante, la discusión desarrollada aún no ha transversalizado los diferentes programas sociales selectivos que hacen uso de IA ni tampoco las iniciativas en materia de asistencia social que abordan a poblaciones en situación de pobreza y pobreza extrema, entre las cuales destaca el Sistema Nacional de Información y Registro Único de Beneficiarios del Estado (SINIRUBE).

En palabras de Cascante y Castro (2024), el SINIRUBE llegó a transformar la política social de carácter asistencial en Costa Rica al implementar la IA tanto para focalizar en determinados grupos poblacionales –bajo la noción de pobreza o pobreza extrema– como para aumentar la tecnificación de los procesos de trabajo de las personas profesionales en trabajo social que históricamente han formado parte de la valoración socioeconómica.

[1] De manera más específica, Jay *et al.* (2024) indican que en Colombia se ha integrado la IA a diversos planes de desarrollo nacional que refieren al impulso de la productividad, la competitividad y la innovación en sectores económicos y sociales.

Con base en lo expuesto, este artículo se plantea las siguientes preguntas. ¿Los marcos normativos, legales y operativos del SINIRUBE se encuentran transversalizados por la ética? ¿Existen dilemas éticos en el cumplimiento de los derechos humanos de las poblaciones registradas en SINIRUBE? ¿Qué desafíos éticos enfrentan las personas profesionales, específicamente en trabajo social, ante la implementación del SINIRUBE en Costa Rica?

Para dar respuesta a las anteriores interrogantes se tomaron en cuenta dos ámbitos de análisis: la discusión aportada específicamente por la *Recomendación sobre la ética de la inteligencia artificial* de la UNESCO (2021) y el *Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina y el Caribe* de la OCDE y la CAF (2022) en aras de posicionar algunas discusiones en torno al sistema. A su vez, el *Código de Ética Profesional* del Colegio de Trabajadores Sociales de Costa Rica (COLTRAS) (2021) brindó una serie de aportes deontológicos, específicamente desde la experiencia de personas profesionales insertas en el Régimen No Contributivo (RNC) de la Caja Costarricense del Seguro Social (CCSS) y el Programa de Apoyos Económicos del Instituto Nacional de Aprendizaje (INA), instituciones que cuentan con convenio vigente con el SINIRUBE desde 2017 (Álvarez, 2021).

A su vez, se colocó como eje central el cuestionamiento planteado por autores como Blázquez (2022), quien cuestiona si la IA puede ser considerada un agente moral –por contar con conciencia– al igual que el ser humano o si el actuar debe centrarse en este último. El desarrollo del estudio de carácter cualitativo involucró la revisión normativa e informes publicados desde el SINIRUBE, el Instituto Mixto de Ayuda Social (IMAS), la Contraloría General de la República (CGR), la CCSS, el INA y el COLTRAS, los cuales posibilitaron comprender la implementación del sistema, los aportes, las limitaciones y las potenciales discusiones éticas. De igual forma, para enriquecer los hallazgos fueron entrevistados 15 profesionales en trabajo social insertos en el Consejo Rector del SINIRUBE, el SINIRUBE propiamente, la CCSS, el INA y el Poder Judicial. Toda esta información fue procesada a partir de matrices y analizada a partir de la técnica de la categorización, la triangulación de datos y la triangulación teórica.



## 1. El SINIRUBE: la incorporación de la IA en la política asistencial costarricense

El Estado costarricense ha transformado históricamente su orientación en materia de inversión social, al pasar de una orientación universalista para la atención de la pobreza propuesta en la década de los años 70 del siglo XX, con la creación de instituciones como el IMAS[2] (Ley Nº 4760 de 1971) y el Fondo de Desarrollo Social

[2] El artículo 2 de la Ley N° 4760 indica: "ARTICULO 2º.- El IMAS tiene como finalidad resolver el problema de la pobreza extrema en el país, para lo cual deberá planear, dirigir, ejecutar y controlar un plan nacional destinado a dicho fin. Para ese objetivo utilizará todos los recursos humanos y económicos que sean puestos a su servicio por los empresarios y trabajadores del país, instituciones del sector público nacionales o extranjeras, organizaciones privadas de toda naturaleza, instituciones religiosas y demás grupos interesados en participar en el Plan Nacional de Lucha contra la Pobreza".

y Asignaciones Familiares (FODESAF)[3] (Ley Nº 5662 de 1974), a una connotación neoliberal de focalización y de mayor tecnificación posterior a la adopción de las medidas indicadas por organismos internacionales (Fondo Monetario Internacional –FMI– y el gobierno de los Estados Unidos) en los llamados Programas de Ajuste Estructural (PAE) durante los años 90.

En la última década del siglo XX, se inició el desarrollo de los llamados Sistemas de Información Social de Beneficiarios en la región. Costa Rica incursionó en sistemas informáticos como el Sistema de Selección de Beneficiarios de la Política Social (SISBEN) y el Sistema de Población Objetivo (SIPO), destinados a priorizar los colectivos poblacionales sujetos de programas de asistencia social en el IMAS (Cascante & Castro, 2024). Al discutirse la relevancia de contar con un sistema unificado que dispusiera de datos para priorizar, administrar y optimizar los fondos públicos destinados a los colectivos poblacionales considerados con "mayor necesidad", se crearon los argumentos y las condiciones para el surgimiento del SINIRUBE a partir de la Ley Nº 9137, publicada en 2013. En esta ley se estableció que dicho sistema funcionaría como órgano desconcentrado adscrito al IMAS (art. 1) bajo los siguientes fines (art. 3):

a. Mantener una base de datos actualizada y de cobertura nacional con la información de todas las personas que requieran servicios, asistencias, subsidios o auxilios económicos, por encontrarse en situaciones de pobreza o necesidad.
b. Eliminar la duplicidad de las acciones interinstitucionales que otorgan beneficios asistenciales y de protección social a las familias en estado de pobreza.
c. Proponer a las instituciones públicas y a los gobiernos locales, que dedican recursos para combatir la pobreza, una metodología única para determinar los niveles de pobreza.
d. Simplificar y reducir el exceso de trámites y requisitos que se les solicita a los potenciales beneficiarios de los programas sociales.
e. Conformar una base de datos que permita establecer un control sobre los programas de ayudas sociales de las diferentes instituciones públicas, con el fin de que la información se fundamente en criterios homogéneos.
f. Disponer de datos oportunos, veraces y precisos, con el fin de destinar de forma eficaz y eficiente los fondos públicos dedicados a los programas sociales.
g. Garantizar que los beneficios lleguen efectivamente a los sectores más pobres de la sociedad, que estos sean concordantes con las necesidades reales de los destinatarios y que las acciones estén orientadas a brindar soluciones integrales y permanentes para los problemas que afectan los sectores de la población más vulnerable.



[3] El artículo 2 de la Ley Nº 5662 indica: "ARTÍCULO 2º.- Son beneficiarios de este Fondo los costarricenses y extranjeros residentes legales del país, así como las personas menores de edad, quienes, a pesar de carecer de una condición migratoria regular en el territorio nacional, se encuentren en situación de pobreza o pobreza extrema, de acuerdo con los requisitos que se establezcan en esta y las demás leyes vigentes y sus reglamentos".

Si bien el artículo 20 de la Ley Nº 9137 dictaminó la articulación con otras instituciones estatales a través de los llamados "convenios de cooperación", recién en 2017 se emitió el Reglamento a la Ley Nº 9137, y en 2019, por medio de la Directriz N° 060-MTSS-MDHIS, se dio paso a la llamada priorización de atención de la pobreza con el uso del SINIRUBE en diferentes instituciones de la Administración Central y descentralizada del sector social.[4] Al analizar los marcos normativos del SINIRUBE, específicamente la Ley Nº 9137 y su reglamento, es posible notar que no se hace referencia específicamente a la IA. Lo más cercano se encuentra en los incisos h) e i) del artículo 8, "Funciones del Consejo Rector, que ordenan:

h. Mantener un sistema de información cruzado, permanente y actualizado, de los sujetos que han tenido acceso a los servicios del Sistema.
i. Resguardar y garantizar la seguridad del Sistema, empleando tecnologías de información, protección y comunicación, con el fin de que las instituciones del Estado cuenten con una información veraz y de probada utilidad

Es decir, en su génesis el SINIRUBE no fue creado bajo la connotación de IA, sino que su incorporación se efectuó posterior a las mejoras consideradas necesarias para cumplir los fines establecidos. Dichas mejoras involucraron, para Acón y Tijerina (2023), que en 2016 se dispusiera de una ventanilla web para consulta de instituciones y una versión que facilitara la interoperación e interconexión entre bases de datos, así como en 2018 se llevara a cabo una serie de mejoras tecnológicas, entre las cuales destaca el apoyo brindado por el Banco Interamericano de Desarrollo (BID) y el Instituto Tecnológico de Massachusetts (MIT) para que se incluyera al SINIRUBE "como parte de un proceso de desarrollo y determinación entre una amplia gama de modelos econométricos y de aprendizaje automático" (Álvarez, 2019, p. 17).



A grandes rasgos, el SINIRUBE se alimenta de bases de datos previamente establecidas provenientes del Tribunal Supremo de Elecciones (TSE), la Dirección General de Migración y Extranjería (DGME), el Registro Nacional de la Propiedad y el Sistema Centralizado de Recaudación (SICERE) de la CCSS. A partir de la información que contiene el Registro de Información Socioeconómica (RIS),[5] y con base en el algoritmo diseñado, se llega a clasificar a los hogares en:

[4] Entre las cuales se menciona el uso obligatorio por parte del Banco Hipotecario de la Vivienda, RNC de la CCSS, el Consejo Nacional de la Persona Adulta Mayor (CONAPAM), el Consejo Nacional de la Persona con Discapacidad (CONAPDIS), el Fondo Nacional de Becas (FONABE), el IMAS, el Programa de Ayudas Económicas y Becas del INA, el Instituto Nacional de las Mujeres (INAMU) y los Programas de Equidad del Ministerio de Educación Pública (MEP), entre otras.

[5] El RIS registra características de la vivienda, acceso a servicios básicos y saneamiento, características del hogar y personas que conforman cada hogar (SINIRUBE, s/f).

- Pobreza extrema
- Pobreza básica
- Vulnerabilidad
- No pobreza
- Por investigar (SINIRUBE, s/f)

Como parte de los convenios firmados, las instituciones participantes cuentan con dos funciones principalmente: el acceso al SINIRUBE a nivel de consulta y la alimentación del Registro Único de Beneficiarios (RUB) con la información de la población usuaria de sus respectivos programas sociales (Álvarez, 2022). Una característica relevante del SINIRUBE que cambió la orientación de la política social de carácter asistencial en el país es la forma de medición unificada de pobreza para todas las instituciones indicadas en la Directriz N° 060-MTSS-MDHIS, ya que, en lugar de hacer uso de un método unidimensional (línea de pobreza) o multidimensional (necesidades básicas insatisfechas, pobreza multidimensional), se pasó a un modelo que, a partir de algoritmos, permite "predecir" la condición de pobreza de un hogar al usar como base la pobreza extrema –categoría de referencia– (Cascante & Castro, 2024).

A diferencia de otros sistemas expertos, con base en lo indicado por Ponce y Torres (2014), el SINIRUBE se considera un sistema simbólico al disponer de un número finito de opciones y reglas. Es decir, según Cascante y Castro (2024), si bien el sistema en mención busca imitar la inteligencia humana para determinar el nivel socioeconómico de los hogares de las personas –"problema" que busca resolver–, no ha alcanzado un nivel de inteligencia automatizada que se alimente de bases de datos –fuera de las previamente establecidas– o aprenda de sus errores.



## 2. El SINIRUBE a la luz de las recomendaciones internacionales sobre la ética de la IA

Aludir a la ética en la IA involucra reflexionar respecto al agente moral y transitar necesariamente por la evolución de los derechos humanos. Siguiendo a Megías (2022), planteamientos iniciales de regulación en la Unión Europea propusieron reconocer personalidad a la IA; sin embargo, en discusiones de la UNESCO se rechazó asignarles personalidad jurídica a dichos sistemas y se enfatizó que la responsabilidad debe recaer en última instancia sobre las personas físicas o jurídicas.

Bajo dicho planteamiento, citando a Coekelberg (2020), Blázquez (2022) señala que las máquinas que intervienen pueden ser agentes instrumentales, pero no agentes morales, ya que por sí mismas carecen de emociones, valores, consciencia e intencionalidad; para el autor, es diferente que los seres humanos deleguen capacidad de actuar a las innovaciones tecnológicas a que eludan responsabilidad

sobre su uso.[6] La particularidad de dicha reflexión en la IA es que –si bien esta es programada por seres humanos para recabar, registrar, combinar y procesar datos– las posibilidades que abren los algoritmos ante el aprendizaje automático desdibujan las fronteras de la comprensión para lo que inicialmente fue programada.

Sobre los derechos humanos –a diferencia de los derechos de la primera (derechos civiles y políticos), segunda (derechos económicos, sociales y culturales) y tercera generación (derechos de la solidaridad o colectivos)–, Flores menciona que la cuarta generación de derechos humanos "viene a responder a nuevas necesidades de la sociedad que no habían aparecido antes, en el contexto de la contaminación de las libertades ante los usos de algunas nuevas tecnologías y avances en las ciencias biomédicas" (2015, p. 34).

Para el caso del sistema estudiado, la discusión ética se encuentra ausente en lineamientos como el Plan Estratégico Institucional-PEI 2019-2023 (SINIRUBE, 2019) y en el más reciente Plan Operativo Institucional 2024 del SINIRUBE (Hernández, 2024). Sin embargo, en su Estrategia Nacional de Inteligencia Artificial de Costa Rica, el MICITT (2024) considera que la IA cuenta con el potencial de llevar a cabo mejoras en la formulación, ejecución y evaluación de las políticas públicas, sobre todo en el acceso a servicios públicos (atención médica y asistencia social), para lo cual es necesario integrar principios éticos en las diferentes fases de aplicación de la IA –etapas iniciales, programación y diseño de algoritmos– (Eje 1: IA ética, segura y responsable) y modernizar la administración pública a través de la mejora en servicios públicos (Eje 4: Gobierno inteligente).



A pesar de lo expuesto, en la misma estrategia y en los proyectos de ley que se encuentran en corriente legislativa, no se halla expuesto de manera explícita el SINIRUBE como sistema que hace uso de la IA, lo cual comporta una evidente invisibilización y una falta de articulación y de armonización entre dicho sistema con los lineamientos políticos nacionales emanados. Al no contarse con referentes éticos en los marcos normativos del SINIRUBE, y al no haber una mención explícita al sistema en la estrategia nacional implementada, la discusión que se expone seguidamente recupera parte de lo expuesto en la *Recomendación sobre la ética de la inteligencia artificial* de la UNESCO (2021) y en el *Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina y el Caribe* de la OCDE y la CAF (2022).

### 2.1. Ética e IA desde la regulación internacional

La discusión respecto a la importancia de regular la IA ha sido fomentada por diferentes organismos internacionales e instancias geopolíticas como la UNESCO, la OCDE y la Unión Europea (Abdala *et al.*, 2019).

[6] A manera de analogía –así como no se puede juzgar el comportamiento de la bala que se dispara e hiere a una persona, o del mismo cuchillo que lastima pero corta el pan que da de comer–, por ende el agente moral respecto a la IA refiere a las diferentes personas que participan en las fases de dichos sistemas, no al sistema mismo.

La UNESCO (2021) adoptó en 2021 un modelo regulatorio a nivel mundial, denominado *Recomendación sobre la ética de la inteligencia artificial*, en el cual enfatiza que las cuestiones éticas asociadas con los sistemas de IA se encuentran presentes en diferentes etapas de sus ciclos de vida: investigación, concepción, desarrollo, funcionamiento, financiación, evaluación, etc. La recomendación insta a los Estados miembros a aplicar –de manera voluntaria– los objetivos indicados en la **Tabla 1**:

**Tabla 1. Objetivos de la *Recomendación sobre la ética de la inteligencia artificial* de la UNESCO**

| |
|---|
| a. Proporcionar un marco universal de valores, principios y acciones para orientar a los Estados en la formulación de sus leyes, políticas u otros instrumentos relativos a la IA, de conformidad con el derecho internacional. |
| b. Orientar las acciones de las personas, los grupos, las comunidades, las instituciones y las empresas del sector privado a fin de asegurar la incorporación de la ética en todas las etapas del ciclo de vida de los sistemas de IA. |
| c. Proteger, promover y respetar los derechos humanos y las libertades fundamentales, la dignidad humana y la igualdad, incluida la igualdad de género; salvaguardar los intereses de las generaciones presentes y futuras; preservar el medio ambiente, la biodiversidad y los ecosistemas; y respetar la diversidad cultural en todas las tapas del ciclo de vida de los sistemas de IA. |
| d. Fomentar el diálogo multidisciplinario y pluralista entre múltiples partes interesadas y la concertación sobre cuestiones éticas relacionadas con los sistemas de IA. |
| e. Promover el acceso equitativo a los avances y los conocimientos en el ámbito de la IA y el aprovechamiento compartido de los beneficios, prestando especial atención a las necesidades y contribuciones de los países de ingreso mediano, incluidos los países menos adelantados (PMA), países en desarrollo sin litoral (PDSL) y pequeños estados insulares en desarrollo (PEID). |

Fuente: UNESCO (2021).

La recomendación de la UNESCO dispone de valores –ideales inspiradores centrados–, principios –primera concretización de dichos valores en exigencias éticas universales– y ámbitos de aplicación –exigencias más concretas– (Megías, 2022). Los valores y principios a los que hace referencia se exponen en la **Tabla 2**:

**Tabla 2. Valores y principios de la *Recomendación sobre la ética de la inteligencia artificial* de la UNESCO**

| Valores | Principios |
|---|---|
| • Respeto, protección y promoción de los derechos humanos, las libertades fundamentales y la dignidad humana<br>• Prosperidad del medio ambiente<br>• y los ecosistemas<br>• Garantizar la diversidad y la inclusión<br>• Vivir en sociedades pacíficas, justas e interconectadas | • Proporcionalidad e inocuidad<br>• Seguridad y protección<br>• Equidad y no discriminación<br>• Sostenibilidad<br>• Derecho a la intimidad y protección de datos<br>• Supervisión y decisión humanas<br>• Transparencia y explicabilidad<br>• Responsabilidad y rendición de cuentas<br>• Sensibilización y educación<br>• Gobernanza y colaboración adaptativas y de múltiples partes interesadas |

Fuente: UNESCO (2021).

Por otra parte, la OCDE y la CAF (2022) cuentan con el documento *Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina y el Caribe*, que no solo recopila estrategias y casos prácticos de IA en la región, sino que plantea que los gobiernos cuentan con la posibilidad de dar prioridad a estrategias nacionales, inversión pública y normativa aplicable asociada con la IA, que permita aumentar la productividad, mejorar la eficiencia y permitir la reducción de los costos. Los principios de la OCDE sobre IA se basan en valores sintetizados en la **Tabla 3**:



**Tabla 3. Principios de la OCDE sobre IA**

| |
|---|
| • La IA debe beneficiar a la población y al planeta mediante el fomento del crecimiento inclusivo, el desarrollo sostenible y el bienestar. |
| • Los sistemas de IA deben diseñarse de manera que respeten el estado de derecho, los derechos humanos, los valores democráticos y la diversidad, y deben incorporar salvaguardias adecuadas –por ejemplo, permitir la intervención humana cuando sea necesario– con miras a garantizar una sociedad justa y equitativa. |
| • Debe haber transparencia y una divulgación responsable de los sistemas de IA para asegurar que las personas entiendan sus resultados y puedan cuestionarlos. |
| • Los sistemas de IA deben funcionar en forma robusta, segura y protegida a lo largo de su ciclo de vida, y los posibles riesgos deben evaluarse y gestionarse en forma permanente. |
| • Las organizaciones y las personas que desarrollen, instalen u operen sistemas de IA deben rendir cuentas de su correcto funcionamiento, de conformidad con los principios previamente mencionados. |

Fuente: OCDE & CAF (2022).

Las recomendaciones que se brindan a los gobiernos nacionales de América Latina y el Caribe para que maximicen los beneficios de la IA en el sector público o minimicen consecuencias negativas o no deseadas se hallan en la **Tabla 4**:

**Tabla 4. Recomendaciones de la OCDE a los gobiernos nacionales de América Latina y el Caribe respecto a IA**

| |
|---|
| 1. Explorar el desarrollo y la ejecución de una estrategia y una hoja de ruta de la IA en el sector público para América Latina y el Caribe a través de un abordaje regional colaborativo. |
| 2. Desarrollar y adoptar estrategias y hojas de ruta nacionales para la IA en el sector público. |
| 3. Elaborar una estrategia de datos nacional para el sector público que abarque distintos aspectos relativos a los datos y que cimente la aplicación de la IA. |
| 4. Explorar las posibilidades de cooperar y colaborar a escala regional para elaborar proyectos e iniciativas de IA en el sector público. |
| 5. Respaldar las actividades que se realicen en materia de IA a nivel subnacional y reflejarlas en políticas e iniciativas de IA más amplias. |
| 6. Fortalecer el énfasis en la implementación de estrategias de IA en el sector público para garantizar la concreción de los compromisos asumidos. |
| 7. Tomar medidas que respalden la sostenibilidad a largo plazo de las estrategias e iniciativas de la IA en el sector público. |
| 8. Llevar a la práctica los Principios de la OCDE sobre Inteligencia Artificial y configurar un marco nacional de ética para una IA fiable. |
| 9. Tener en cuenta como un elemento central las consideraciones sobre el uso de una IA fiable en el sector público que se identifican en el presente informe. |
| 10. Arbitrar los medios para generar una capacidad de liderazgo sostenida a nivel central e institucional que guíe y supervise la adopción de la IA en el sector público. |
| 11. Valerse de técnicas de gobernanza anticipada de la innovación para prepararse para el futuro. |
| 12. Tener en cuenta como un elemento central las consideraciones sobre gobernanza que se identifican en el presente informe. |
| 13. Tener en cuenta enfáticamente como un elemento central los habilitadores críticos de la IA en el sector público que se identifican en el presente informe. |

Fuente: OCDE & CAF (2022).

En torno a la ética, el planteamiento expuesto lleva a cabo un énfasis en los datos al considerar que, a pesar de que los sistemas modernos de IA se han basado en ellos, "[…] la disponibilidad, calidad, integridad y pertinencia de los datos no son suficientes para garantizar la imparcialidad y la inclusividad de las políticas y las decisiones, o

para reforzar su legitimidad y la confianza pública" (OCDE & CAF, 2022, p. 64). Bajo esta línea, los principios de buenas prácticas asociados con la ética de los datos en el sector público (OCDE, 2021, citada en OCDE & CAF, 2022) refieren a:

- Gestionar los datos con integridad.
- Conocer y observar, en todos los niveles de gobierno, las disposiciones pertinentes acerca del acceso confiable a los datos, su intercambio y uso.
- Incorporar consideraciones éticas sobre los datos a los procesos de toma de decisiones gubernamentales, organizacionales y del sector público.
- Monitorear y retener el control sobre los datos que se ingresan, en especial los destinados a fundamentar el desarrollo y entrenamiento de los sistemas de IA, y adoptar un enfoque de riesgo respecto de la automatización de las decisiones.
- Ser específico con relación al propósito para el cual se usan los datos, en particular en el caso de los datos personales.
- Definir límites de acceso, intercambio y empleo de los datos.
- Ser claro, inclusivo y abierto.
- Publicar datos abiertos y el código fuente.
- Ampliar el control de personas y colectivos sobre sus datos.
- Rendir cuentas y ser proactivo en la gestión de riesgos.



En general, la regulación expuesta permite contar con un marco normativo internacional y orientador común que ha de ser tomado en cuenta por Costa Rica al ser país integrante de la UNESCO desde 1992 y de la OCDE desde 2021.[7] Tanto la *Recomendación sobre la ética de la inteligencia artificial* de la UNESCO (2021) como el *Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina y el Caribe* de la OCDE y la CAF (2022) carecen de una conceptualización propia de la ética. Empero, ambos lineamientos coinciden en la relevancia de incorporar la IA en conjunto con los derechos humanos y la ética.

El énfasis de la recomendación de la UNESCO (2021) se orienta a la sociedad en general (personas, grupos, comunidades, instituciones y empresas del sector privado), mientras que la OCDE y la CAF (2022) establece como público meta de su documento a los gobiernos de América Latina y el Caribe. Con base en lo indicado por Megías (2022), la UNESCO (2021) reconoce que lo planteado corresponde a recomendaciones, puesto que existen diversas orientaciones éticas y culturas que

[7] Costa Rica es miembro de la UNESCO desde 1992 según lo indicado en el Reglamento de la Comisión Costarricense de Cooperación con la UNESCO N° 20984-RE-E-C-CYT-RNEM y más recientemente se reiteró su adhesión en 2007 con lo indicado en el Reglamento de Organización y Funcionamiento de la Comisión Costarricense de Cooperación con la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) N° 34.276. También se adhirió como miembro de la OCDE en 2021 con base en lo indicado en el Acuerdo sobre los términos de la adhesión a la Convención de la Organización para la cooperación y el desarrollo económico, suscrita en San José, París y el Protocolo Adicional N° 1 y Nº 2 a la Convención de la organización para la Cooperación y Desarrollo Nº 9981. Por encontrarse Costa Rica fuera del rango geográfico y direccionamiento político emanado por la Unión Europea, en este artículo no se incorporan los aportes de dicho modelo regulatorio.

impactan en los valores éticos predominantes a nivel local y regional, lo cual no abre paso necesariamente a la posibilidad de vulnerar los derechos humanos ni las libertades fundamentales. En este punto, la OCDE y la CAF (2022) coinciden en que los gobiernos promuevan sistemas de IA que respeten tanto los derechos humanos como los valores democráticos.

### 2.2. Alcances y riesgos del SINIRUBE en Costa Rica

Llevar a cabo un análisis ético del SINIRUBE conllevó, en primera instancia, recuperar los informes efectuados por instancias a nivel nacional e internacional que, por un lado, reconocen de manera positiva en su discurso los alcances que ha generado el uso del sistema en la política social de carácter asistencial, mientras que, por el otro, involucró destacar los riesgos en su operacionalización que –para efectos de la discusión que interesa– potencialmente se convierten en dilemas éticos.

En el ámbito nacional, el Programa Estado de la Nación (2023) recalcó que el SINIRUBE "tiene un rol clave como herramienta de asignación, gestión, seguimiento y planificación de la política social selectiva que se dirige a segmentos específicos de la población" (2023, p. 46), ya que ha permitido innovar en el diseño de políticas con base en "evidencias" y mejorar la eficiencia en la ejecución de fondos públicos.

A nivel internacional, el SINIRUBE recibió en 2020 el premio a la innovación en gestión pública efectiva de la Organización de Estados Americanos (OEA) en la categoría de Innovación en Gobierno Inteligente (Álvarez, 2022). Por su parte, la OCDE (2023) refirió que gracias al SINIRUBE se han mejorado las políticas sociales en el país, al eliminar duplicidades y aumentar la cobertura de identificación de potenciales personas beneficiarias: "El SINIRUBE debe convertirse en la piedra angular de las políticas sociales y ser la herramienta central para seleccionar a los beneficiarios de todos los programas sociales. Esto ayudaría a mejorar la focalización y evaluación de las políticas sociales" (OCDE, 2023, p. 58).



Empero, el primer dilema ético necesario de hacer referencia es la focalización, ya que, si bien en Costa Rica la asistencia social no es un derecho (Vásquez, 2014), existe un discurso meritocrático constante y de ataque a las poblaciones que, excluidas de otras formas para reproducir sus condiciones de vida, hacen uso de ella. En otras palabras, existe un juicio moral que abarca la política social desde los años 90 del siglo XX, en el cual para el Estado costarricense el ser sujeto de asistencia social "sin merecerlo" llega a convertirse en un "pecado".

Independientemente de los posicionamientos políticos a favor del sistema, y sobre todo de la lectura estatal que considera de manera positiva la focalización y regulación de la inversión social,[8] los aportes expuestos principalmente por la

[8] No se hace uso del término "gasto social" pues se considera que reproduce una lógica mercantil y no una basada en los derechos humanos.

UNESCO (2021) brindan un amplio marco de entendimiento asociado con los riesgos del diseño y la implementación de la IA en iniciativas de carácter público. Con base en dichos aportes es posible puntualizar una serie de dilemas éticos para el caso del SINIRUBE en Costa Rica:

- Acceso y privacidad de los datos
- Opacidad y caja negra
- Desinformación
- Prejuicios y estereotipos
- Deshumanización del proceso

Los dilemas éticos a los que se hace referencia prestan atención en las personas profesionales que participaron en el diseño, la fiscalización, la promoción y el mantenimiento del sistema, considerados los primeros agentes morales; es decir, los dilemas transversalizan a los profesionales que laboran directamente en el sistema y a quienes forman parte del Consejo Rector. Asimismo, debido a que el actuar del SINIRUBE genera consecuencias para el acceso a los programas sociales que priorizan la atención de la pobreza en colectivos poblacionales, se contempla dentro del análisis a las personas solicitantes a quienes se les aprueba el apoyo económico y a aquellos a quienes les es rechazada, en el marco del cumplimiento de sus derechos humanos.



En relación con el acceso y privacidad de los datos, es pertinente reconocer en primera instancia que la OCDE y la CAF (2022) han propuesto dentro de sus buenas prácticas el monitorear y retener el control sobre ellos, así como la UNESCO dentro de sus principios se refiere al derecho a la intimidad y protección de datos:

> "La privacidad, que constituye un derecho esencial para la protección de la dignidad, la autonomía y la capacidad de actuar de los seres humanos, debe ser respetada, protegida y promovida a lo largo del ciclo de vida de los sistemas de IA. Es importante que los datos para los sistemas de IA se recopilen, utilicen, compartan, archiven y supriman de forma coherente con el derecho internacional y acorde con los valores y principios enunciados en la presente Recomendación, respetando al mismo tiempo los marcos jurídicos nacionales, regionales e internacionales pertinentes" (UNESCO, 2021, p. 22).

En este caso el SINIRUBE –tal como se mencionó– se alimenta de información brindada por diferentes bases de datos instituciones públicas (TSE, DGME, SICERE) y de los reportes de población usuaria facilitados por las entidades que cuentan con convenio de cooperación, en el marco de la obligatoriedad de la Directriz N° 060-MTSS-MDHIS. Con base en lo indicado por la Agencia de Protección de Datos

de los Habitantes (PRODHAB), citada por Acón y Tijerina (2023), la información que dispone el sistema no se encuentra sujeta al consentimiento informado, al encontrarse respaldada por lo indicado en la Ley N° 8969 de Protección de la Persona frente al tratamiento de sus datos personales, que indica: "Artículo 8.- Excepciones a la autodeterminación informativa del ciudadano […] e) La adecuada prestación de servicios públicos. f) La eficaz actividad ordinaria de la Administración, por parte de las autoridades oficiales". Bajo el mismo marco normativo, la opción de consulta pública –que en algún momento contó el SINIRUBE– fue deshabilitada, así como según el perfil de ingreso las personas profesionales cuentan con acceso a diferente información (SINIRUBE, s/f).[9]

Si bien para Blázquez existe una "paradoja de la transparencia", en tanto dentro de la IA "el acceso, procesamiento de datos masivos y su respectiva combinación, permiten acceder –de forma invasiva– a través de las aplicaciones a información personal de carácter privado" (Blázquez, 2022, p. 265), el SINIRUBE ha creado mecanismos que regulan el acceso a la información y protección de los datos, así como no hay posibilidad de consulta a datos personales por entidades privadas que dispongan de la información para fines lucrativos o generar desinformación, como ha alertado Megías (2022).

No obstante, con base en los límites de acceso, intercambio y empleo de datos indicados por la OCDE y la CAF (2022), existe un riesgo, y por ende un dilema ético, en el acceso a información sensible que el Consejo Rector del SINIRUBE pueda facilitar a profesionales consideradas dentro del umbral de "otras" en las ciencias sociales, a quienes en el estudio socioeconómico se les brinde acceso por necesidad de la institución y no por contar con competencias en su formación básica para analizar las situaciones que se reportan (como puede ser la violencia doméstica).



Si la información disponible en SINIRUBE es tomada para decisiones estatales –como se indica en los incisos a) y e) del artículo 3 de la Ley N° 9137–, se debe prever que no en todas las ocasiones los datos de la población registrada se encuentran actualizados, puesto que algunas de las personas entrevistadas reportaron que "el problema para responder porque no siempre está actualizado […] Porque a veces hay personas que parecen con registros del 2020, del 2021, del 2019" (persona 3, comunicación personal, octubre de 2023).

En otra línea, la opacidad y caja negra del SINIRUBE se enmarca en el principio de transparencia y explicabilidad de la UNESCO, que alude a:

[9] Los perfiles de ingreso según el *Manual: Consulta Registro de Información Socioeconómica RIS en SINIRUBE* refieren a: visualización básica (contiene información general de la persona consultada, no permite visualizar clasificación de pobreza), análisis inicial (destinado a personas profesionales que no cuentan con un perfil en ciencias sociales, permite acceder a la conformación del hogar y clasificación emitida por SINIRUBE), estudio económico (profesionales en ciencias sociales u otra rama que ejecutan análisis de los hogares, no cuentan con acceso a información considerada sensible) y estudio socioeconómico (profesionales en ciencias sociales: trabajo social, psicología y orientación, etc., que efectúan un análisis de los hogares y asignan los llamados "beneficios económicos"; cuentan con acceso a las diferentes situaciones reportadas) (SINIRUBE, s/f).

> "La transparencia y la explicabilidad de los sistemas de IA suelen ser condiciones previas fundamentales para garantizar el respeto, la protección y la promoción de los derechos humanos, las libertades fundamentales y los principios éticos. La transparencia es necesaria para que los regímenes nacionales e internacionales pertinentes en materia de responsabilidad funcionen eficazmente. La falta de transparencia también podría mermar la posibilidad de impugnar eficazmente las decisiones basadas en resultados producidos por los sistemas de IA y, por lo tanto, podría vulnerar el derecho a un juicio imparcial y a un recurso efectivo, y limita los ámbitos en los que estos sistemas pueden utilizarse legalmente" (UNESCO, 2021, p. 23).

La opacidad no solo refiere al tratamiento de los datos –mencionados en el anterior riesgo–, sino que deviene del secretismo que involucró el diseño del algoritmo, una supuesta caja negra que carece de capacidad explicativa y dificultad para su inteligibilidad, incluso para las personas que saben del tema (Prodnik, 2022). Lo cual dota de un poder absoluto sobre las resoluciones que llega a emitir:

> "De hecho, desconocemos por completo su funcionamiento, pero la experiencia permite constatar que cuando el algoritmo adopta una resolución, esta se convierte en una especie de veredicto –concesión o denegación de préstamos, hipotecas, puestos de trabajo, ayudas sociales etc.,– ante el cual no es posible objetar ni recurrir, a pesar de que esa decisión haya podido alimentarse de datos parciales, eventualmente incorrectos o insuficientes, o que han podido ser malinterpretados, sin revelar nada a cambio" (Blázquez, 2022, p. 267).



Tal como se mencionó, el algoritmo utilizado por el SINIRUBE fue creado por el MIT bajo la lógica de un modelo econométrico predictivo de la pobreza (Cascante & Castro, 2024), para el cual –tomando lo indicado por Abdala *et al.* (2019)– no sabe cómo se procesa la información según las bases de datos de las cuales se alimenta, lo que puede derivar en resultados sesgados y de alguna manera discriminatorios.

Dicho desconocimiento o desinformación respecto a cómo funciona el algoritmo utilizado por SINIRUBE incumple tanto el principio de transparencia y explicabilidad como el de sensibilización y educación, planteado por la UNESCO (2021), puesto que es una constante en las personas profesionales entrevistadas. Si bien varias personas consultadas indican que al inicio de su implementación en sus respectivos espacios institucionales recibieron capacitación asociada con cómo consultar y alimentar el sistema –según el rol asignado–, no existe claridad de las interrelaciones que desarrolla el algoritmo para emitir su criterio:

> "[…] para mí es un misterio, porque teniendo casos con dos condiciones muy similares, muy similares de vida, limitaciones económicas, tenencia de bienes, composición de grupo familiar, un grupo familiar me sale en condición de pobreza básica y el otro grupo me sale como vulnerable. Entonces hay un indicador, alguno de todos, que no está coincidiendo en los dos y los raros, que para mí los dos están exactamente en las mismas condiciones […] Hemos encontrado casos en donde la persona cuenta con una red de apoyo fuertísima para efectos de la atención de sus necesidades, pero SINIRUBE la coloca en condición de riesgo, entonces puede presentar la solicitud y va a tener derecho al beneficio; y hemos encontrado casos de personas en condición de riesgo social en donde tienen o requieren realmente el acompañamiento, pero SINIRUBE las califica como no pobres o vulnerables" (persona 11, comunicación personal, noviembre de 2023).

De la mano con la opacidad y caja negra deviene el riesgo de incurrir en potenciales prejuicios y estereotipos. En general, diversos sistemas de IA analizados por autores como Blázquez (2022) y Megías (2022) se han caracterizado por reproducir una lectura de los grupos poblacionales en los que se privilegian determinadas formas de pensamiento y se excluyen a quienes no coincidan con sus planteamientos. Aunque los informes recopilados en torno al SINIRUBE no han mostrado evidencia empírica para comprender potenciales prejuicios y estereotipos, ante la opacidad del sistema y el desconocimiento respecto al funcionamiento de su algoritmo es inevitable considerar la posibilidad de que esto ocurra. Los prejuicios y estereotipos sobre los que se alerta de manera ética al SINIRUBE refieren necesariamente a la información obtenida en las bases de datos de las que se alimenta, las cuales cuentan con referencia a aspectos como el género, la edad y la condición migratoria que se desconoce si están incidiendo en la comprensión de la IA respecto a la medición de pobreza:



> "[…] la obscuridad que preside los tratamientos de datos y la subsiguiente combinatoria correlacional permite a sus creadores mantener en secreto el diseño del algoritmo así como los pasos que se han dado, pudiendo aducir, eventualmente, el derecho a la propiedad intelectual. El problema adicional que puede derivarse es que, la aplicación de los sistemas de inteligencia artificial, pueden generar después, o incluso perpetuar, prejuicios o estereotipos establecidos, desfavoreciendo más aun a grupos étnicos o sociales que se han visto históricamente marginados" (Blázquez, 2022, p. 266).

Por último, se considera que existe un riesgo en el SINIRUBE relacionado con la deshumanización del proceso. El primero de los valores al que hace referencia la

UNESCO (2021) es el respeto, protección y promoción de los derechos humanos, las libertades fundamentales y la dignidad humana; un valor que "se complementa con el de diversidad e inclusión, que se concreta en descartar el uso de la IA para *homogeneizar a los seres humanos* e impedir el desarrollo personal de acuerdo con lo que nos diferencia a unos de otros" (Megías, 2022, p. 146).[10]

En este sentido, no solo el Estado costarricense ha reproducido nuevamente un discurso meritocrático que se enfoca en encontrar "a los más pobres de los pobres" a través del SINIRUBE (Cascante & Castro, 2024), sino que ha llegado a implementar un método único de medición de pobreza que homologa múltiples situaciones de exclusión y vulnerabilidad que enfrenta la población, incongruente con el mismo inciso g) de la Ley N° 9137. Un claro ejemplo de ello es la eliminación de la medición de "necesidades especiales" estipulada en el Reglamento del RNC –Reforma de sesión Nº 9112, 20 de julio de 2020 de la Junta Directiva de la CCSS–, medición de la línea de pobreza de manera ampliada que le permitía a las personas profesionales en trabajo social de la Gerencia de Pensiones de la CCSS tomar en cuenta gastos derivados por motivos de enfermedad o discapacidad.[11]

Bajo esta misma línea, es menester hacer una diferenciación: si bien el SINIRUBE posibilita que las personas catalogadas en situación de pobreza o pobreza extrema accedan de manera más efectiva a los programas asistenciales (persona 2, comunicación personal, noviembre de 2023) –al cumplir el más relevante criterio de focalización establecido–, para aquellas poblaciones que quedan excluidas la situación genera un *impasse*. Dicho punto es notable cuando a la población no beneficiaria se le da una resolución del SINIRUBE que refiere a no encontrarse en pobreza o pobreza extrema y se le encomienda modificar su información directamente en alguna oficina del IMAS para que el caso pueda volver a valorarse, lo cual se complejiza al ser el SINIRUBE un órgano desconcentrado del IMAS que carece de oficinas para la atención del público:



> "La reforma consiste en trasladar la responsabilidad de evaluar la condición de pobreza de los solicitantes de un beneficio de aseguramiento, tarea que anteriormente realizaba la Caja Costarricense de Seguro Social, al Instituto Mixto de Ayuda Social. Esto se lleva a cabo mediante el Sistema Nacional de Información y Registro Único de Beneficiarios del Estado (SINIRUBE). A tres años de producida esta reforma, preocupa a la Defensoría el importante número de denuncias que se continúan conociendo ante lo que se considera una dilación del IMAS en la aplicación de los estudios socioeconómicos, lo cual es comprensible tomando en cuenta la

[10] La cursiva es del autor.
[11] Esta situación es coincidente con la medición de la canasta derivada de la discapacidad que se encuentra estipulada en el artículo 15 de la Ley N° 9379 para la Promoción de la Autonomía Personal de las Personas con Discapacidad, la cual no se está aplicando debido a la medición de pobreza implementada por el SINIRUBE.

> labor propia que al respecto ya el IMAS debe atender en todo el país" (Defensoría de los Habitantes, 2023, p. 39).

Con base en lo expuesto, la deshumanización del proceso involucra pasar por alto las barreras que colectivos vulnerables pueden enfrentar,[12] no solo para trasladarse físicamente, sino también para entender por qué su situación no puede ser atendida directamente por una persona profesional de la institución en la cual realizó la solicitud.[13] Al articular la deshumanización con la focalización mencionada inicialmente, existe un vacío respecto al cumplimiento de la eliminación de las duplicidades de las acciones interinstitucionales a partir del SINIRUBE –indicado en el inciso b) de la Ley N° 9137–, ya que el sistema reporta como duplicidad cualquier beneficio económico recibido en un mismo núcleo familiar independientemente de las necesidades o finalidades que busca atender (persona 1, comunicación personal, noviembre de 2023). Es decir, no existe claridad de cómo se operativiza dicho fin, en tanto los programas estatales cuentan con diferentes objetos de intervención y, por ende, ante diferentes finalidades no se puede asumir que existe una duplicidad en la asignación de transferencias monetarias.

### 3. El SINIRUBE desde la ética deontológica de trabajo social

Los códigos de ética refieren a una serie de principios, normas y preceptos que permiten regular el comportamiento humano profesional bajo el estandarte de fundamentos morales y éticos de carácter general en la cotidianidad de las personas profesionales (Castillo, 2010).



Para el caso del análisis ético propuesto el gremio de profesionales en trabajo social dispone del Código de Ética Profesional que refiere, en su artículo 1º, al "conjunto de principios, normas, valores y deberes éticos y morales que rigen el ejercicio profesional en trabajo social" (COLTRAS, 2021, p. 6).[14] Los valores y principios que reúne la ética deben guiar la conducta personal del gremio de profesionales, sobre todo en pro del respeto de los derechos humanos, la justicia, la autonomía, el bien común y la dignidad de las personas (COLTRAS, 2021). En relación con el SINIRUBE, las personas profesionales en trabajo social se convierten en agentes moralizadores responsables de ejecutar la orientación focalizadora del Estado costarricense, para lo cual entra en discusión las posibilidades de hacer frente ante la imposición del criterio de la IA por encima de la postura profesional.

---

[12] Según el artículo 6 del Reglamento de RNC, la población beneficiaria se encuentra compuesta por personas adultas mayores, personas "inválidas", viudas en desamparo, huérfanos e "indigentes". Cabe mencionar que se hace uso de la terminología establecida en el reglamento de RNC a pesar de considerarse que parte de ella es peyorativa e incongruente con la normativa asociada con los derechos humanos.

[13] A las denuncias interpuestas ante la Defensoría de los Habitantes se suman otras que se han presentado ante el Juzgado de Seguridad Social del Poder Judicial contra la CCSS por motivo de rechazo de RNC (persona 15, comunicación personal, octubre de 2024).

[14] Otros marcos normativos vigentes para el gremio de profesionales en trabajo social, tal como la Ley Nº 3943 Orgánica del Colegio de Trabajadores Sociales o su respectivo Reglamento, no refieren al interés de la presente discusión.

### 3.1. El criterio humano subalterno a la IA

El Código de Ética Profesional de las personas profesionales en trabajo social de Costa Rica menciona la digitalización de documentos y tecnología de la siguiente manera:

**Tabla 5. Mención de la digitalización de documentos y uso de la tecnología en el Código de Ética profesional de trabajo social**

| Capítulo | Artículos |
|---|---|
| IV. Derechos y deberes relacionados con las personas | Artículo 28º – Informes sociales y otros registros escritos asociados a las atribuciones profesionales. Las personas profesionales en trabajo social: |
| X. Uso de tecnología y sistemas de información en el ejercicio profesional | Artículo 46º – Incorporación de tecnologías y sistemas de información en el quehacer profesional. Las personas profesionales en trabajo social:<br>a. Se actualizarán permanentemente en el uso y manejo de nuevas herramientas tecnológicas y sistemas de información *que coadyuven al mejoramiento del desempeño profesional, docente e investigativo* y procurarán innovar las formas de intervención y atención de sus objetos y sujetos de trabajo.<br>b. Darán a conocer y promoverán en la población el uso de medios tecnológicos e informativos que les permitan desarrollar conocimientos, habilidades, capacidades y prácticas para defender y ampliar sus derechos y propiciar el mejoramiento de sus condiciones de vida. |
| X. Uso de tecnología y sistemas de información en el ejercicio profesional | Artículo 47º – Precauciones en el uso de tecnologías y sistemas de información. Las personas profesionales en trabajo social:<br>a. Reconocerán que el uso de la tecnología, los sistemas de información digital y las redes sociales representan un reto para el mantenimiento de los principios éticos relacionados con la privacidad, confidencialidad, conflictos de interés, competencia y resguardo de la documentación. Las personas profesionales en trabajo social tratarán de desarrollar las habilidades y obtener los conocimientos necesarios para evitar la comisión de acciones contrarias a la ética.<br>b. La incorporación de las tecnologías y sistemas de información digital en el quehacer profesional no deben conducir a la deshumanización, afectación, disminución, ni desvalorización del ejercicio y el criterio profesional, los espacios laborales y las competencias propias del gremio de trabajo social. Por el contrario, las personas profesionales deben permanecer vigilantes ante cualquier política, medida o directriz que pueda perjudicar al gremio y, especialmente, a la población cuando no estén en riesgo de enfrentar una tecnocratización indebida. |
| X. Uso de tecnología y sistemas de información en el ejercicio profesional | Artículo 48º – Protección de documentos y registros digitales. Las personas profesionales en trabajo social deben utilizar medios tecnológicos adecuados para proteger y salvaguardara los documentos de trabajo, los informes sociales y los registros que aparezcan en formato digital y contengan información privada y confidencial sobre la población intervenida profesionalmente. |

Fuente: elaboración propia a partir de COLTRAS (2021). Nota: las cursivas son del autor.

En el marco de su amplio entendimiento de tecnología, el Código de Ética Profesional permite ubicar el uso de SINIRUBE, por lo que pueden acontecer escenarios "en donde las obligaciones éticas de las personas profesionales en trabajo social entren en conflicto con las políticas del centro laboral o con las leyes y regulaciones nacionales" (COLTRAS, 2021, p. 3).

Un primer elemento en el que se considera que el SINIRUBE riñe con el código refiere al artículo 2º, que promueve entre el colectivo de profesionales en trabajo social el respeto de la dignidad de las personas basadas en los derechos humanos, ya que no hay un real reconocimiento de la diversidad humana al promoverse métodos hegemónicos de medición de pobreza que han dejado de lado las particularidades de determinados grupos poblacionales. El método prospectivo de medición de la pobreza que utiliza el SINIRUBE (Cascante & Castro, 2024), en el que se privilegia a través de algoritmos la noción de pobreza extrema, no es coherente con otras mediciones existentes en el país, tal como la canasta derivada de la discapacidad que implementaba RNC de la CCSS. A su vez, con base en el inciso c del artículo 3º, el colectivo profesional debe defender el acceso de las personas a obtener los recursos que satisfagan sus necesidades humanas, un reto que se complejiza ante la focalización y tecnificación de los procesos de trabajo en los cuales –para el caso de RNC de la CCSS– se carece de posibilidades actuales para hacer frente al criterio del SINIRUBE.



Parafraseando a Cascante y Castro (2024), la decisión de considerar la clasificación de los hogares como criterio determinante no es un aspecto que se encuentra explícito en los marcos normativos existentes (Ley Nº 9137, Reglamento a la Ley Nº 9137, Directriz N° 060-MTSS-MDHIS o Directriz N° 081-MTSS-MDHIS), sino que ha sido resultado de decisiones políticas de actores tomadores de decisiones como la Junta Directiva de la CCSS en las reformas realizadas a la normativa interna, tal como el Reglamento de RNC, lo cual es coincidente con el principio de buena práctica de la OCDE y la CAF (2022), que indica que la ética de datos en el sector público debe adoptar un enfoque de riesgo asociado con la automatización de las decisiones.

En su principio de proporcionalidad e inocuidad, la UNESCO destaca que, además de utilizarse y justificarse el uso de la IA para lograr un objetivo determinado, y que el método utilizado no vulnere los valores fundamentales reconocidos en su recomendación, se debe contemplar que:

> "En los casos en que se entienda que las decisiones tienen un impacto irreversible o difícil de revertir o que puedan implicar decisiones de vida o muerte, la decisión final debería ser adoptada por un ser humano. En particular, los sistemas de IA no deberían utilizarse con fines de calificación social o vigilancia masiva" (UNESCO, 2021, p. 20).

Esto se complementa con el principio de supervisión y decisiones humanas, en el que se insta a los Estados miembros a velar que en toda etapa de los sistemas de IA exista tanto una supervisión individual como pública (UNESCO, 2021). Con base en Megías (2022), de este principio deriva la obligación de velar por que toda decisión final se encuentre bajo control humano y que sean las personas físicas o jurídicas las responsables de rendir cuentas, en tanto también "la IA debe estar al servicio del hombre [ser humano] y no este último al servicio de aquella. La IA debe ser una herramienta de optimización del trabajo humano y no una expresión de reemplazo de las tareas humanas" (Figueroa, 2024, p. 154).

Por otra parte, desde la promoción de la coherencia profesional, en el inciso a y b del artículo 10º, el SINIRUBE no promueve el mayor beneficio de la población. Como sistema, alude a alimentarse de diferentes fuentes de información, pero la falta de datos actualizados y la imposición del criterio no son coherentes con evitar los procesos de revictimización (artículo 14º), puesto que la información no es fidedigna y en su lugar se están tomando decisiones en ocasiones sobre datos desactualizados. Los datos disponibles en SINIRUBE se encuentran mediados por la PRODHAB; empero no hay posibilidad de resguardo de datos contenidos en la documentación digital que es creada –como promueve el inciso g del artículo 20 de secreto profesional, confidencialidad y privacidad–, ya que los datos no fueron confiados directamente por la población a los profesionales en trabajo social.



Aquellas personas profesionales que se encuentren en puestos tomadores de decisiones (jefaturas, supervisiones o jerarquías organizacionales) tienen, según el inciso i del artículo 36º, la obligación de "i. Abogar dentro y fuera de sus instituciones o ámbitos de trabajo por los recursos necesarios para el acceso, ejercicio pleno de los derechos de la población y el adecuado desarrollo del ejercicio laboral de trabajadores sociales bajo sus órdenes" (COLTRAS, 2021, p. 34). Sin embargo, para efectos de llevar a cabo la evaluación e investigación social (artículo 38º), los profesionales en trabajo social de RNC han sido cooptados de hacer uso de las técnicas y los métodos considerados más idóneos para investigar los fenómenos objetos de interés –tal como indica el inciso e) del artículo mencionado–, ya que no existe alternativa para hacer frente a la medición de pobreza establecida por el SINIRUBE, así como carecen de libertad en el ejercicio profesional (inciso a) del artículo 41º) para elegir o aplicar otras técnicas que consideren idóneas.

En general, los dilemas éticos que atraviesan los profesionales en trabajo social a partir de la implementación del SINIRUBE en determinados espacios profesionales –no en todos aplica por igual– evidencian una racionalización instrumental que afecta las relaciones sociales y reproduce una lógica tecnológica o una ideología algorítmica –en palabras de Prodnik (2022)–, así como también una fascinación "bigdataísta" en la que el avance tanto de la robótica como de la IA se idolatra sin cuestionamientos, independientemente de la afectación que genere (Goñi, 2019).

## Consideraciones finales

En concordancia con Megías (2022), quien enfatiza que la regulación de los sistemas de IA debe ser tanto jurídica como ética, es posible notar que a nivel jurídico el SINIRUBE de Costa Rica cuenta con un marco base que se inició en 2013 con la creación del sistema (Ley Nº 9137), el cual paulatinamente se fue robusteciendo hasta ser reglamentado en 2017 y alcanzar en 2019 una priorización de su uso en diferentes instituciones del apartado estatal, según la Directriz N° 060-MTSS-MDHIS y Directriz N° 081-MTSS-MDHIS. Empero la misma situación no acontece con una orientación ética marcada, ya que –a pesar de disponerse de orientaciones a nivel internacional como la *Recomendación sobre la ética de la inteligencia artificial* de la UNESCO (2021), el *Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina* de la OCDE y la CAF (2022), y la Estrategia Nacional de Inteligencia Artificial de Costa Rica a cargo del MICITT (Álvarez, 2023), junto con otros proyectos de ley en corriente legislativa–, no existe una armonización explícita con el SINIRUBE. Tal como se mencionó, la orientación inicial del SINIRUBE en la Ley N° 9137 no hizo mención del uso de IA, sino que, a la luz de las transformaciones históricas, fue incorporada con apoyo del MIT en 2018. Sin embargo, la incorporación postergada no es una justificación para que actualmente se carezca de marcos éticos claros que conlleven a la reflexión de las diferentes personas que fungen como agentes morales.

La ética en el presente análisis es posicionada como una oportunidad para visualizar los riesgos y dilemas de la IA que no permiten olvidar la responsabilidad del agente moral. Dicha noción de agente moral, planteado por Blázquez (2022), es un punto base de referencia, ya que, a pesar de los avances en automatización independientes de los procesos que simulan el razonamiento humano, reiteramos que la IA carece de conciencia propia e intencionalidad en su actuar, por lo que, como instrumento o máquina, traslada la discusión de la ética a aquellas personas que participan en sus fases de diseño, mejoras, fiscalización e implementación. En su particularidad, resulta interesante que los riesgos y su concreción en los dilemas éticos del SINIRUBE –a diferencia de la focalización– son coincidentes con las situaciones expuestas por otros autores cuando mencionan a la IA en general (Blázquez, 2022; Figueroa, 2024; Megías, 2022; Prodnik, 2022), los cuales evidencian una mayor afectación en el cumplimiento de los derechos humanos de la población que queda excluida de los servicios asistenciales, al no encontrarse ubicadas en pobreza o pobreza extrema por el sistema, y en el criterio de personas profesionales en trabajo social, que ha quedado relegado de manera subalterna a lo dictado por el SINIRUBE.

Si bien desde los años 80 el Estado costarricense ha marcado un rumbo hacia un proyecto neoliberal que abrió paso a la focalización y a una paulatina tecnificación de los procesos en materia de asistencia social (Cascante & Castro, 2024), los marcos de entendimiento basados en los derechos humanos son el estandarte para que la ética se encuentre en la palestra de discusión. Incluso es posible decir que, si no es con la ética, no será posible estrechar el distanciamiento y la polarización que se está llevando a cabo entre el actuar de la IA y el del ser humano.

**Financiamiento**

El estudio que posibilitó la elaboración del presente artículo fue financiado por la Vicerrectoría de Investigación de la Universidad de Costa Rica. A su vez, el artículo forma parte de los resultados del proyecto de investigación "Pry01-233-2022-Transformaciones del rol de Trabajo Social a partir de la implementación del SINIRUBE en el Instituto Nacional de Aprendizaje y Régimen No Contributivo de la Caja Costarricense del Seguro Social en la Región de Occidente", desarrollado durante 2023 y 2024 en la Sede de Occidente de la Universidad de Costa Rica.

**Bibliografía**

Abdala, M. B., Lacroix Eussler, S. & Soubie, S. (2019). La política de la Inteligencia Artificial: sus usos en el sector público y sus implicancias regulatorias. Documento de Trabajo N° 185. Buenos Aires: CIPPEC. Recuperado de: https://www.cippec.org/publicacion/la-politica-de-la-inteligencia-artificial-sus-usos-en-el-sector-publico-y-sus-implicancias-regulatorias/.

Acón, K. & Tijerina, L. (2023). El SINIRUBE: habilitador de política social de precisión en Costa Rica. Banco Interamericano de Desarrollo. Recuperado de: https://publications.iadb.org/es/el-sinirube-habilitador-de-politica-social-de-precision-en-costa-rica.



Álvarez, E. (2019). Informe de labores del 2018 al I trimestre 2019. San José de Costa Rica: SINIRUBE. Recuperado de: https://www.imas.go.cr/sites/default/files/docs/SINIRUBE-159-%20Informe%20de%20labores%20a%20marzo%202019.pdf

Álvarez, E. (2021). Informe de labores IV Trimestre 2020. San José de Costa Rica: SINIRUBE. Recuperado de: https://www.sinirube.go.cr/wp-content/uploads/2023/07/Informe-de-labores-IV-Trimestre-2020.pdf.

Álvarez, E. (2022). Informe de Gestión 2016-2022. San José de Costa Rica: SINIRUBE. Recuperado de: https://www.imas.go.cr/sites/default/files/docs/Informe%20de%20Gesti%C3%B3n%20periodo%202015-2022.pdf.

Álvarez, M. (2023). Costa Rica será el primer país de Centroamérica en tener una estrategia de Inteligencia Artificial. UNESCO. Recuperado de: https://www.unesco.org/es/articles/costa-rica-sera-el-primer-pais-de-centroamerica-en-tener-una-estrategia-de-inteligencia-artificial.

Blázquez, J. (2022). La paradoja de la transparencia en la IA: opacidad y explicabilidad. Atribución de responsabilidad. Revista Internacional de Pensamiento Político, 17, 261-272. DOI: https://doi.org/10.46661/revintpensampolit.7526.

Cascante, R. & Castro, F. (2024). El SINIRUBE: entre la focalización y tecnificación de la Política Social Asistencial costarricense. En M. Sánchez (Comp), Retroceso de la política social: su incidencia en la vida de la población y en el trabajo profesional de Trabajo Social (127-160). San José de Costa Rica: Colegio de Trabajadores Sociales de Costa Rica. Recuperado de: https://trabajosocial.or.cr/retroceso-de-la-politica-social-su-incidencia-en-la-vida-cotidiana-y-en-el-trabajo-profesional-de-trabajo-social/.

Castillo, C. (2010). Fundamentos de los códigos de ética de los colegios profesionales. Revista Educación, 34(1), 119-141. Recuperado de: https://revistas.ucr.ac.cr/index.php/educacion/article/view/501.

Castro, V., Hernández, J., Morales, M., Morera, O. & Obando, J. (2023). Proyecto de ley: Ley de regulación de la Inteligencia Artificial en Costa Rica. San José de Costa Rica: Asamblea Legislativa de la República de Costa Rica. Recuperado de: https://d1qqtien6gys07.cloudfront.net/wp-content/uploads/2023/05/23771.pdf.

COLTRAS (2021). Código de Ética Profesional. COLTRAS. Recuperado de: https://trabajosocial.or.cr/wp-content/uploads/2024/04/codigo-etica-profesional-_-2.pdf.

Defensoría de los Habitantes (2023). Informe anual 2022-2023. San José de Costa Rica: Defensoría de los Habitantes. Recuperado de: https://www.dhr.go.cr/images/informes-anuales/if2022_2023.pdf.



Figueroa, E. (2024). Inteligencia Artificial (AI) emergente: ¿riesgo potencial para los Derechos Humanos? Ius Inkarri - Revista de la Facultad de Derecho y Ciencia Política, 13(15), 143-167. DOI: https://doi.org/10.59885/iusinkarri.2024.v13n15.07.

Flores, L. (2015). Temas actuales de los derechos humanos de última generación. México: Benemérita Universidad Autónoma de Puebla. Recuperado de: https://archivos.juridicas.unam.mx/www/bjv/libros/9/4304/13.pdf.

Goñi, J. (2019). Innovaciones tecnológicas, inteligencia artificial y derechos humanos en el trabajo. Documentación laboral, (117), 57-72. Recuperado de: https://dialnet.unirioja.es/servlet/articulo?codigo=7095888.

Gutiérrez, J. (2023). Lineamientos para el uso de inteligencia artificial en contextos universitarios. GIGAPP Estudios Working Papers, 10(272), 416-434. Recuperado de: https://www.gigapp.org/ewp/index.php/GIGAPP-EWP/article/view/331.

Hernández, G. (2024). Plan Operativo Institucional 2024. San José de Costa Rica: SINIRUBE. Recuperado de: https://www.sinirube.go.cr/media/2024/02/POI-SINIRUBE-2024-1.pdf.

Izquierdo, O. (2023). Proyecto de ley: Ley para la promoción responsable de la Inteligencia Artificial en Costa Rica. San José de Costa Rica: Asamblea Legislativa

de la República de Costa Rica. Recuperado de: https://d1qqtien6gys07.cloudfront.net/wp-content/uploads/2023/09/23919.pdf.

Jay, W., Padilla, M. & Rodelo, M. (2024). Políticas públicas ante la Revolución de la inteligencia artificial en Colombia. Revista Venezolana de Gerencia, (106), 865-883. DOI: https://doi.org/10.52080/rvgluz.29.106.26.

MICITT (2024). Estrategia Nacional de Inteligencia Artificial de Costa Rica. San José de Costa Rica: MICITT. Recuperado de: https://www.micitt.go.cr/sites/default/files/2024-10/Estrategia%20Nacional%20de%20Inteligencia%20Artificial%20de%20Costa%20Rica%20ESP.pdf.

Megías, J. (2022). Derechos Humanos e Inteligencia Artificial. Revista Dikaiosyne, (37), 139-163. Recuperado de: https://www.saber.ula.ve/bitstream/handle/123456789/47846/articulo6.pdf?sequence=1&isAllowed=y.

Obando, J. (2024). Proyecto de ley: Ley para la implementación de sistemas de Inteligencia Artificial (IA). San José de Costa Rica: Asamblea Legislativa de la República de Costa Rica. Recuperado de: https://d1qqtien6gys07.cloudfront.net/wp-content/uploads/2024/08/24484.pdf.

OCDE (2023). Estudios Económicos de la OCDE: Costa Rica 2023. París: OCDE. Recuperado de: https://www.comex.go.cr/media/9642/estudios-econ%C3%B3micos-de-la-ocde-costa-rica-survey-spanish_master_230130_1744_230131_1346pdfx.pdf.

OCDE (2024). Recommendation of the Council on OECD Legal Instruments Artificial Intelligence. OECD/LEGAL/0449. París: OECD Legal instruments. Recuperado de: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

OCDE & CAF (2022). Uso estratégico y responsable de la inteligencia artificial en el sector público de América Latina y el Caribe. Estudios de la OCDE sobre Gobernanza Pública. París, Francia: OECD Publishing. DOI: https://doi.org/10.1787/5b189cb4-es.

Ponce, J. & Torres, A. (2014). 1. Introducción y Antecedentes de la Inteligencia Artificial. En: A. Casali, J. Hernández, E. Martínez, F. Ornelas, O. Pedreño, J. Ponce, F. Quezada, E. Scheihing, A. Silva, A. Torres, M. Torres, Y. Túpac, N. Vakhnia & C. Zavala (Comps.), Inteligencia Artificial. Argentina: Proyecto LATIn. Recuperado de: https://rephip.unr.edu.ar/items/8beb7e66-bc8b-48e1-bd7c-3dbc3dfcceb9.

Prodnik, J. (2022). La lógica algorítmica del capitalismo digital. Revista Hipertextos, 10 (18). Recuperado de: https://revistas.unlp.edu.ar/hipertextos/article/view/14603/13660.

Programa Estado de la Nación (2023). Informe Estado de la Nación 2023. San José de Costa Rica: Consejo Nacional de Rectores (CONARE). Recuperado de: https://estadonacion.or.cr/?informes=informe-estado-de-la-nacion-2023.

SINIRUBE (s/f). Manual: Consulta Registro de Información Socioeconómica RIS en SINIRUBE. San José de Costa Rica: SINIRUBE. Recuperado de: https://aprendamos.sinirube.go.cr/pluginfile.php/3135/mod_folder/content/0/Manual%20-%20Consulta%20de%20RIS.pdf.

SINIRUBE (2019). Plan Estratégico Institucional-PEI 2019-2023. San José de Costa Rica: SINIRUBE. Recuperado de: https://www.sinirube.go.cr/media/2023/07/PEI-Completo_Sinirube.pdf.

UNESCO (2021). Recomendación sobre la ética de la inteligencia artificial. París: UNESCO. Recuperado de: https://www.unesco.org/es/articles/recomendacion-sobre-la-etica-de-la-inteligencia-artificial.

Vásquez, A. (2014). Asistencia social. Un estudio de los principales referentes teórico-metodológicos. Emancipação, 14(2), 261-276. Recuperado de: https://dialnet.unirioja.es/servlet/articulo?codigo=5403074.

# Los dilemas de la inteligencia artificial. De los sistemas sociotécnicos a la ética del diseño

# Os dilemas da inteligência artificial. Dos sistemas sociotécnicos à ética do design

# *The Dilemmas of Artificial Intelligence. From Socio-Technical Systems to Design Ethics*

**Santiago José Roca Petitjean**
**y Alejandro Elías Ochoa Arias** *



Este artículo explora el debate acerca de la ética en el desarrollo de inteligencia artificial (IA), con el propósito de contribuir con la comprensión de sistemas sociotécnicos que integren la deliberación ética y la codificación de criterios de decisión en el desarrollo de sistemas inteligentes. Se inicia con un examen de los temas de ética y responsabilidad en el contexto de los sistemas inteligentes y los agentes autónomos. A continuación, se consideran los marcos interpretativos de bienes públicos (ciencia abierta) y bienes privados (ciencia mercantil), abordando los discursos de dos entes multilaterales y dos empresas, respectivamente. En este artículo se considera la coconstrucción de sistemas sociotécnicos donde se reconozca la pertinencia de la ética y la responsabilidad, se conciba el desarrollo informático como práctica social y las demandas éticas sean transferidas operativamente a los sistemas inteligentes a través de métodos de diseño orientado a valores. Finalmente, se discute la relación entre la ética y el desarrollo tecnológico con respecto a la cocreación de sistemas sociotécnicos y sistemas inteligentes que cumplan con criterios de responsabilidad.

**Palabras clave:** ética; responsabilidad; sistema sociotécnico; inteligencia artificial; diseño orientado a valores

* *Santiago José Roca Petitjean:* investigador del Centro Nacional de Desarrollo e Investigación en Tecnologías Libres (CENDITEL), Venezuela. Correo electrónico: sroca@cenditel.gob.ve. ORCID: https://orcid.org/0000-0002-3701-3409. *Alejandro Elías Ochoa Arias:* profesor de la Universidad Austral de Chile (UACh), Chile. Correo electrónico: alejandro.ochoa@uach.cl. ORCID: https://orcid.org/0000-0001-8464-7108.

Este artigo explora o debate sobre a ética no desenvolvimento da inteligência artificial (IA), com o objetivo de contribuir para a compreensão de sistemas sociotécnicos que integram a deliberação ética e a codificação de critérios de decisão no desenvolvimento de sistemas inteligentes. Começa com um exame das questões de ética e responsabilidade no contexto de sistemas inteligentes e agentes autônomos. A seguir, são considerados os marcos interpretativos dos bens públicos (ciência aberta) e dos bens privados (ciência comercial), abordando os discursos de duas entidades multilaterais e de duas empresas, respectivamente. Este artigo considera a coconstrução de sistemas sociotécnicos onde a relevância da ética e da responsabilidade é reconhecida, o desenvolvimento computacional é concebido como uma prática social e as demandas éticas são transferidas operacionalmente para sistemas inteligentes através de métodos de design orientado para valor. Por fim, discute-se a relação entre ética e desenvolvimento tecnológico no que diz respeito à cocriação de sistemas sociotécnicos e sistemas inteligentes que atendam a critérios de responsabilidade.

**Palavras-chave:** ética; responsabilidade; sistema sociotécnico; inteligência artificial; design orientado para valor

*This article explores the debate about ethics in the development of artificial intelligence (AI), with the purpose of contributing to the understanding of sociotechnical systems that integrate ethical deliberation and the codification of decision criteria in the development of intelligent systems. It begins with an examination of the issues of ethics and responsibility in the context of intelligent systems and autonomous agents. Next, the interpretative frameworks of public goods (open science) and private goods (commercial science) are considered, addressing the discourses of two multilateral entities and two companies, respectively. This article considers the co-construction of sociotechnical systems where the relevance of ethics and responsibility is recognized, computer development is conceived as a social practice and ethical demands are operationally transferred to intelligent systems through value-oriented design methods. Finally, the relationship between ethics and technological development is discussed with respect to the co-creation of sociotechnical systems and intelligent systems that meet responsibility criteria.*

***Keywords:*** *ethics; responsibility; socio-technical system; artificial intelligence; value-oriented design*

## Introducción

La inteligencia artificial (IA) se ha convertido en uno de los exponentes más significativos de la revolución digital. Esta premisa equivale a afirmar su importancia como parte de una revolución tecnológica, orientada a transformar los sistemas productivos y organizacionales, y por tanto a fortalecer el paradigma tecnoeconómico propio de las sociedades digitalizadas (Pérez, 2020). En otras palabras, la IA es, al mismo tiempo, una innovación parcial, parte de un conjunto de sistemas tecnológicos consolidados y motor de cambio socioeconómico y cultural. Éste se expresa en infinidad de propuestas académicas, empresariales, legislativas y gubernamentales que giran en torno a la necesidad de adecuar los procesos sociales y las organizaciones para la implementación de IA, explotar sus beneficios y asumir sus costos. Y en paralelo, se prepara la conformación de sistemas sociotécnicos mediados por la intervención de sistemas inteligentes y de agentes autónomos, en áreas tan estratégicas como las comunicaciones, la educación y el empleo, sin olvidar los sectores militar y médico que poseen características críticas relativas a la responsabilidad.

Las innovaciones tecnológicas tienden a transformar las relaciones socioproductivas, generando el abandono de las formas organizacionales precedentes y presionando para la conformación de nuevas relaciones (Pérez, 2020). Pero, al mismo tiempo, tales cambios preceden a las experiencias de aprendizaje cultural y de adaptación institucional a las nuevas condiciones productivas. Esto significa que la revolución digital contrae una mora con respecto a la generación de propuestas éticas, legislativas y políticas que permitan reducir los aspectos nocivos y aprovechar los beneficios de la innovación. Y también implica que, en general, se carece de un marco sociocrítico para interpretar y responder a las transformaciones tecnológicas. En tal sentido, es posible indagar en la tradición cultural en busca de propuestas para afrontar el impacto de las revoluciones tecnoeconómicas.



La ética puede considerarse la piedra angular de la filosofía política, el derecho y las políticas públicas, dado que en sus bases subyacen cuestiones sobre la naturaleza del bien, la importancia de las normas sociales o el significado de la vida plena. No obstante, hoy día la racionalidad instrumental y la racionalidad estratégica parecen dar forma a la interacción entre las personas y su entorno social, obstaculizando la deliberación propia de una racionalidad comunicativa, que debería apuntar a fomentar un consenso pluralista sobre esas cuestiones. En otras palabras, la consolidación de la separación entre el "sistema" y "el mundo de vida", de acuerdo con Habermas (1999), se ha materializado en una colonización profunda del primero sobre el segundo. Por lo tanto, tras las narrativas tecnooptimistas y tecnopesimistas, se encuentra una suerte de determinismo tecnológico que induce a pensar que los problemas sociales merecen una respuesta mediada por la tecnología (o incluso, que los problemas socioeconómicos son consecuencia de la aplicación deficiente de soluciones instrumentales), y las posibilidades de generar dinámicas de deliberación comunitaria que tengan influencia en el espacio público (por encima de las lógicas de mercado) parecen extrañas y más aún, con escasa legitimidad.

Como consecuencia, la humanidad se encuentra frente a los retos que le impone la presente revolución tecnológica, con pocas herramientas para la conformación de un marco sociocrítico que permita comprender el lugar de la IA ante las exigencias que plantean materias como la ética, el derecho, la filosofía, la economía política y la gestión pública. De esa manera, es necesario abordar dos temas: cómo se construyen las alternativas éticas para convertirse en criterios de decisión, y de qué manera pueden ser transferidos éstos a los efectos de la producción tecnológico-cultural, como en el caso del desarrollo responsable de IA. Con el fin de abordar estas cuestiones, este trabajo plantea que la tecnología debe concebirse como un sistema sociotécnico, donde tienen relevancia los valores, intereses y relaciones de los agentes tecnológicos y aquellos que quedan al margen. De ese modo, la IA forma parte de "un sistema sociotécnico donde las nuevas tecnologías interactúan con elementos sociales" (Buijsman, Klenk & van den Hoven, 2025, p. 60). Tal enfoque permite no solo observar las relaciones sociotécnicas desde una perspectiva crítica, sino también ser creativos en la formulación de alternativas de respuesta a los problemas sociales desde la preocupación por la innovación abierta y responsable (Terrones, 2020).

Como parte de los sistemas sociotécnicos, la ética del desarrollo tecnológico se presenta como una exigencia que se realiza a los entes de regulación y de mercado en escenarios de conflicto. Si bien puede partirse de la tradición histórica de la ética, también es necesario reconocer las particularidades de la situación actual, con miras a señalar las condiciones que facilitan el reclamo de manejo ético de la tecnología y de aproximarse a las alternativas de prácticas de desarrollo éticas y responsables. De este modo, es necesario explorar cómo se construyen las opciones éticas en el caso de la IA, para entonces encontrar opciones teórico-metodológicas que permitan "abordar los desafíos de la ética de la IA combinando teorías éticas [...] con un enfoque de diseño" (Buijsman, Klenk & van den Hoven, 2025, pp. 76-77). Sin embargo, la preocupación por los valores y principios que deben guiar el desarrollo responsable de IA también debe contribuir con la fundamentación de políticas globales, iniciativas jurídicas, esquemas de diseño tecnológico y prácticas de mercado congruentes entre sí, que favorezcan una tendencia hacia la democratización de las dinámicas de apropiación social de las tecnologías.



En definitiva, este artículo se propone explorar la importancia de la ética en el desarrollo de la IA, con miras a ofrecer insumos para el diseño de sistemas sociotécnicos donde tengan cabida tanto la deliberación en torno a problemas fundamentales como la cocreación de bienes tecnológicos. En primer lugar, se parte de una exploración de temas como la ética y la responsabilidad frente al desarrollo de sistemas inteligentes y agentes autónomos. Posteriormente, se realiza una revisión de algunos discursos en torno a la ética y la responsabilidad en la implementación de IA, particularmente desde los contextos interpretativos de los bienes públicos (ciencia abierta) y privados (ciencia mercantil). Finalmente, se ofrece una discusión que se plantea interpretar la vinculación entre ética y producción tecnológica a la luz de la construcción de alternativas éticas, institucionales y de diseño pertinentes para la conformación de sistemas sociotécnicos de carácter más ético y responsable.

## 1. Ética y responsabilidad en la IA

Para abordar el tema es necesario tomar en cuenta diferentes enfoques éticos y su aplicación en el desarrollo e implementación de sistemas inteligentes. Puede plantearse que existen tres enfoques éticos conocidos: ética de la virtud, que se preocupa de los rasgos de carácter que deben cultivarse para tener una buena vida; consecuencialismo, interesado por las formas de comportamiento que contribuyen con el bienestar; y deontología, que se ocupa de normas universales de comportamiento (Buijsman, Klenk & van den Hoven, 2025). Respectivamente, la ética se origina en la concepción del individuo, en la percepción de los resultados o en la aceptación de normas consideradas universales. Estos enfoques pueden servir como guía para plantearse cuestiones clásicas de la ética, cómo por ejemplo qué es el bien, pero también para inferir distintos criterios de decisión, como si para alcanzar el bien deben respetarse ciertas normas o si deben buscarse resultados que contribuyan a crear condiciones de vida más deseables.

En ese sentido, tales enfoques pueden ofrecer información sobre qué es ético, cómo comportarse de forma ética, y en qué sentido los efectos de las acciones deben ser congruentes con lo anterior. En otras palabras, forman parte de los fundamentos de una ética aplicada, que debe contribuir a esclarecer las cualidades del desarrollo tecnológico al mismo tiempo que establecer un anclaje con contextos más amplios. Así, con respecto al diseño de sistemas, la cuestión podría enunciarse de la siguiente manera:



> "¿Qué virtudes se deberían inculcar a quienes desarrollan y utilizan la IA? ¿Cómo puede contribuir la educación de los ingenieros a esto, para inculcar virtudes fundamentales como la conciencia del contexto social de la tecnología y el compromiso con el bien público y la sensibilidad hacia las necesidades de los demás?" (Buijsman, Klenk & van den Hoven, 2025, p. 71).

Estas preocupaciones han tomado vigencia recientemente con los avances de la IA, que posee mayores capacidades de actuación que otras tecnologías, pero que, del mismo modo, no posee agencia moral ni puede hacerse responsable por los resultados de su desempeño, lo que plantea la necesidad de establecer de forma sistemática principios éticos y de responsabilidad adecuados para su implementación (Buijsman, Klenk & van den Hoven, 2025). Por ejemplo, en el caso de los agentes autónomos, se ha planteado que éstos pueden encarar tres tipos de moralidad: operacional, propia de los diseñadores y usuarios; funcional, de los sistemas que pueden realizar cálculos morales; y genuina, en el caso de máquinas que puedan crear sus propios sistemas morales de decisión (Cortina, 2024). Por lo tanto, resulta necesario partir de los aspectos de ética y responsabilidad de los seres humanos, en cuanto que la reflexión sobre las implicaciones éticas de la tecnología tiene como eje al ser humano, sus interpretaciones e interacciones, incorporadas en sistemas sociotécnicos de carácter histórico y complejo (Coeckelbergh, 2021).

Resulta claro que cualquier marco ético que se intente insertar en un sistema inteligente es el resultado de la deliberación entre personas. En otras palabras, los conceptos éticos susceptibles de ser programados en las máquinas se derivan de la tradición histórica de reflexión sobre los valores éticos. Por lo tanto, no es sorprendente que la ética aplicable a los sistemas inteligentes pueda ser interpretada como una proyección de enfoques éticos en los procesos de producción de sistemas tecnológicos. Floridi (2023) realiza un planteamiento representativo del estado de la cuestión. Por una parte, reconoce los aportes de diferentes organizaciones en la elaboración de principios éticos para la IA, los cuales reúnen unos 47 axiomas. Posteriormente, los pasa a través del tamiz de la bioética para luego ofrecer una propuesta que incluye a los conjuntos mencionados:

- *Beneficencia:* mejorar la calidad de vida, respetar la dignidad humana y cuidar el medioambiente.
- *No maleficencia:* evitar causar daño, proteger la privacidad y la seguridad, vigilar las capacidades de la tecnología.
- *Autonomía:* empoderar a las personas para que tomen sus propias decisiones de manera libre y consciente.
- *Justicia:* fomentar la igualdad de oportunidades, mantener la cohesión social y prevenir el trato injusto.
- *Explicabilidad:* asegurar que los sistemas y sus decisiones sean comprensibles y que los responsables rindan cuentas por ellos.



El procedimiento de basarse en principios ha resultado fructífero para que diferentes organizaciones, públicas y privadas, ofrezcan sus propias perspectivas sobre los valores rectores en la implementación de IA, con diferente grado de abstracción (de las normas generales a las prácticas de desarrollo) y distinta orientación (de las normas abstractas a los estándares de mercado). Claro está, cualquier ente del ecosistema digital puede presentar sus propios juegos de principios. Por ejemplo, OpenAI (s/f) cuenta con una carta fundacional que plantea que sus objetivos son: beneficios ampliamente distribuidos; seguridad a largo plazo; liderazgo técnico y orientación cooperativa. En otro documento, se responde a la cuestión del desarrollo de IA responsable de la siguiente manera:

> "El desarrollo responsable de la IA implica tomar medidas para garantizar que los sistemas de IA tengan un riesgo aceptablemente bajo de dañar a sus usuarios o a la sociedad e, idealmente, aumentar su probabilidad de ser socialmente beneficiosos. [...] Diremos que un sistema de IA se ha desarrollado de manera responsable si los riesgos de que cause daños están en niveles que la mayoría de la gente consideraría tolerables, teniendo en cuenta su gravedad, y que la cantidad de evidencia que fundamenta estas estimaciones de riesgo también se consideraría aceptable" (Askell, Brundage & Hadfield, 2019, p. 2).

Desde esta perspectiva consecuencialista, el desarrollo responsable de IA consiste en promover sistemas beneficiosos para la sociedad, mientras que resulten "aceptablemente" no dañinos. Un sistema inteligente ha sido desarrollado de forma responsable si el riesgo de que cause daño resulta tolerable para la sociedad. En este sentido, la ética del desarrollo es evaluada por la opinión pública (que puede expresarse de innumerables formas), y colocará en una balanza sus riesgos y sus beneficios. Sin embargo, esta aproximación plantea la cuestión de si un comportamiento puede considerarse ético solo por ser considerado beneficioso (o inocuo) por la opinión general o, en otras palabras, si existen diferencias entre la ética, la opinión de la mayoría y lo que se considera "verdadero" desde una posición de poder. Claro está, es necesario desgranar diferentes posiciones para aproximarse a una comprensión de lo que puede entenderse como "responsabilidad" en cada contexto, y entonces procurar un diálogo entre las diferentes propuestas.

La responsabilidad puede abordarse como un caso especial de la ética. La facultad de rendir cuenta de las motivaciones, realización y efecto de las acciones debe entenderse en tres sentidos: como responsabilidad moral (énfasis en las intenciones de los agentes), responsabilidad causal (énfasis en las acciones) y responsabilidad de rol (énfasis en el lugar organizacional) (Lauwaert & Oimann, 2025). Esos tipos de responsabilidad pueden estar relacionados. Por ejemplo, una persona que causa un accidente de forma no intencional tiene una responsabilidad causal, pero no moral. En cambio, alguien que comete una acción maliciosa de forma intencional tiene responsabilidad moral y causal. La combinación de los tres tipos de responsabilidad puede dar lugar a diferentes formas de juicio en distintos contextos de acción, y se vuelve más compleja cuando se analizan escenarios donde agentes artificiales pueden ser causantes de accidentes o acciones maliciosas.



En este sentido, algunos autores opinan que la agencia de los sistemas inteligentes crea una "brecha de responsabilidad", en cuanto que la responsabilidad es un atributo de los sujetos conscientes y no de los artefactos (Lauwaert & Oimann, 2025). En ese caso, si una máquina ocasiona un accidente, no solo es imposible imputar culpa, sino que tampoco se pueden señalar a otros responsables. Desde otra perspectiva, que no se pueda atribuir responsabilidad moral a una máquina no significa que no existan responsables. Entre otras razones, en los sistemas sociotécnicos participan agentes como reguladores, financistas, diseñadores, desarrolladores y usuarios, por lo que pueden considerarse diferentes modos de responsabilidad, distribuida entre los roles vinculados con la implementación de un artefacto. En ese sentido, el argumento de la brecha de responsabilidad está limitado como pretexto para liberar sistemas inteligentes sin supervisión, como se explica a continuación:

> "La brecha de responsabilidad no significa que los desarrolladores y usuarios de sistemas de IA no tengan obligaciones especiales. Por el contrario, dicha tecnología afirma precisamente la importancia de los deberes morales. Debido a que el poder de toma de decisiones se está transfiriendo a esa tecnología, y que a menudo es imposible

> predecir exactamente qué decisión tomará, los desarrolladores de sistemas de IA deben pensar incluso más cuidadosamente que otros diseñadores de tecnología sobre los posibles efectos indeseables que pueden resultar de las decisiones de los algoritmos" (Lauwaert & Oimann, 2025, pp. 104-105).

En este contexto, la falta de control directo sobre las acciones de los sistemas inteligentes no puede constituirse en favor de la ausencia de responsabilidad. En cambio, la declaración de responsabilidad refuerza la necesidad de implementar estándares de desarrollo que permitan mejorar aspectos como las normas de regulación y de calidad, entre muchos otros. En tal sentido, la implementación de sistemas inteligentes se encuentra subordinada al respeto de la dignidad humana y de los derechos fundamentales, aspectos susceptibles de ser reconocidos por las instituciones, pero también como parte de las dinámicas de innovación (Terrones, 2020).

Se considera entonces que los problemas de ética y responsabilidad tienen su origen en los contextos sociales (Coeckelbergh, 2021), dado que, en sentido estricto, es el único dominio donde un conflicto se presenta como tal. Esto permitirá observar dichos problemas como relaciones entre valores y sistemas sociotécnicos y, de esa forma, reconocer que hacen parte de marcos de diseño y evaluación de aplicaciones. En cuanto que la IA se considera imbricada en un sistema sociotécnico, la admisión de la ética y la responsabilidad en su desarrollo también "implica cambios en el sistema sociotécnico en el que está integrada" (Buijsman, Klenk & van den Hoven, 2025, p. 67). Por lo tanto, es necesario examinar la relación recursiva entre los dispositivos y sus contextos. La consideración sobre la ética de la IA se amplía, entonces, para abarcar la caracterización ética de sus contextos de diseño e implementación.



Los temas de la ética y la responsabilidad no escapan al interés de gobiernos y organizaciones multilaterales. Por ejemplo, el *Global Index on Responsible AI* (Adams *et al.*, 2024), considera 19 áreas temáticas agrupadas en tres dimensiones: derechos humanos e IA, gobernanza responsable de IA y capacidades de IA responsable. Cada área temática abarca tres pilares diferentes del ecosistema de IA: marcos gubernamentales, acciones gubernamentales e iniciativas de actores no estatales. Entre sus principales conclusiones, señala que las estrategias nacionales carecen de principios éticos y no son vinculantes; que las medidas para proteger los derechos humanos en el contexto de IA aún son limitadas; y que existen aspectos preocupantes en torno a temas como la igualdad de género, la inclusión y la equidad. También señala que existen brechas en materia de seguridad y confiabilidad de la IA, así como en protección de los trabajadores. Finalmente, recomienda incrementar la participación de las universidades y la sociedad civil, así como la cooperación internacional, en el cumplimiento de estándares de IA responsable.

El *Artificial Intelligence Index Report 2024* (AI Index Steering Committee, 2024) identifica varias categorías relevantes en materia de inteligencia artificial responsable:

privacidad, gestión de datos, transparencia, capacidad de explicación, justicia, seguridad y protección. Este informe ofrece información técnica sobre numerosos incidentes en materia de responsabilidad de IA. Por ejemplo, señala que desde 2013 los incidentes se han multiplicado por 20 (aunque pueden estar subreportados), y que existen muchos aspectos a mejorar en esta materia, como el manejo de datos privados de los usuarios (que en algunos casos pueden ser extraídos de las aplicaciones de IA generativa) y la utilización de material protegido con derechos de autor en el entrenamiento de IA. Los dos reportes mencionados son avalados por diferentes entidades multilaterales, gobiernos, organizaciones no gubernamentales y empresas, lo que permite afirmar que existe un nivel de conciencia significativo sobre los retos de la ética y la responsabilidad de la IA.

Ahora bien, la reflexión acerca de las relaciones entre la ética y tecnologías como la IA posee varias aristas. Puede tratarse sobre la ética de las superinteligencias al estilo del poshumanismo; del marco ético que adoptarán los seres humanos para interactuar con sistemas inteligentes o de la programación de máquinas que actúen de acuerdo con principios éticos programados (Cortina, 2024). Tales enfoques pueden tener afinidades y diferencias. Además, existen obstáculos comunes, como por ejemplo la dificultad de instrumentar los principios éticos en forma de criterios de decisión en los sistemas de información (Canca, 2020). Por lo tanto, la ética puede hacerse presente en los sistemas inteligentes en forma de los esquemas de valores de sus desarrolladores, en las características funcionales de los programas y en los efectos de su implementación en diferentes escenarios.



Se han vislumbrado tres enfoques disponibles para esta discusión, y por tanto susceptibles de ser tomados en cuenta en la conformación de sistemas sociotécnicos y el diseño de sistemas inteligentes (Cortina, 2025; Misselhorn, 2022):

1. *Arriba a abajo:* con insumos de la tradición deontológica (énfasis en las normas) y utilitarista (énfasis en las consecuencias). Se enfoca en la formulación e inserción de principios morales en forma de capacidades funcionales de las máquinas.
2. *Abajo a arriba:* se propone establecer marcos interpretativos basados en los contextos y en los sujetos, como la ética de la virtud. Se basa en una ética orientada por valores relevantes para cada situación, por lo que se propone encontrar patrones y adaptarlos al aprendizaje de las máquinas.
3. *Híbrido:* como combinación entre ambos enfoques, se parte de un marco de principios que podrían ser adaptados inductivamente a contextos específicos. Se trataría de transferir valores éticos definidos colaborativamente a las características funcionales de los sistemas.

En este contexto, resulta necesario establecer referentes éticos que permitan que los seres humanos actúen de forma creativa y consensuada de acuerdo con determinadas reglas, para que, dado el caso, las mismas se transfieran al diseño y programación

de la IA. Esta es la importancia del enfoque híbrido, el cual "opera sobre la base de un marco predefinido de valores morales que luego se adapta a contextos específicos mediante procesos de aprendizaje" (Misselhorn, 2022, p. 41). Por ejemplo, se necesitaría establecer, a través de un proceso fundado en una ética comunicativa, un conjunto de normas válidas en cada tipo de actividad (basadas, por ejemplo, en su bien interno) para entonces derivar reglas operativas y programarlas en los sistemas inteligentes. Tal sería la premisa de un enfoque dialógico basado en la hermenéutica crítica (Cortina, 2024), que puede favorecer la participación cívica en el desarrollo e implementación de IA.

Uno de los referentes de este enfoque es la ética de la virtud, cuya comprensión se puede hacer con referencia a obras de autores como MacIntyre (2013). Este autor plantea que las virtudes son importantes para la realización de prácticas sociales, lo que implica comprender la naturaleza de sus bienes, identificar las virtudes más importantes y realizar acciones pertinentes para alcanzarlas. El enfoque puede sintetizarse en el concepto de "práctica virtuosa", comprendida como una forma de actividad humana establecida socialmente, que permite desarrollar sus bienes inherentes al mismo tiempo que alcanzar mayores niveles de perfección. Por lo tanto, una práctica es virtuosa si se plantea la búsqueda de un bien interno y tiene en miras el cultivo de la excelencia. Por ejemplo, la producción de *software*, como ejercicio instrumental, no es necesariamente virtuosa en sí misma, pero el desarrollo responsable requiere la incorporación de virtudes que conduzcan a la obtención de mejores resultados, así como a una mayor comprensión de su carácter como práctica virtuosa. Esta es una manera de concebirlo como una actividad orientada a fines éticos, y por lo tanto a la generación de bienes sociales. En otras palabras, la búsqueda virtuosa del bien interno de una práctica debe facilitar el logro de sus bienes externos. Una práctica de desarrollo ética debería ofrecer respaldo a sistemas sociotécnicos más éticos.



Si se toma como referente la teoría de las prácticas virtuosas, sería necesario que los agentes sociales definan los bienes internos de las diferentes prácticas, para que los especialistas logren "incorporar en las programaciones las normas que deben seguir las máquinas para ayudar a alcanzar el bien interno" (Cortina, 2025, p. 120). En el primer momento reside el carácter dialógico de la propuesta. La deliberación en torno a valores clave puede contribuir a establecer un marco de referencia en la esfera pública, que brinde el contexto de interpretación a entes reguladores, empresas y usuarios para orientar el desarrollo e implementación de IA. El segundo momento, la codificación de decisiones éticas, no tiene carácter meramente instrumental, sino que, basados en el perfil constructivo de los sistemas sociotécnicos, se puede aspirar a consolidar una idea de tecnología que implique una forma distinta de plantearse los medios, fines, recursos y agentes que toman parte en el desarrollo tecnológico. Ambos momentos pueden contribuir a fomentar la democratización de los procesos de innovación desde una perspectiva cívica.

Puesto que la deliberación en torno a valores éticos debe tener como resultado efectivo el surgimiento de una modalidad de desarrollo responsable, el tema del diseño

orientado a valores juega un papel fundamental (Van den Hoven, Vermaas & Van de Poel, 2015). Desde la perspectiva que se ha venido discutiendo, se podría "diseñar un marco, contando con teorías éticas acreditadas y, dentro de él, articular las normas y las reglas adecuadas para el razonamiento, la decisión y la actuación, obtenidas a partir de las particularidades y las experiencias concretas de los afectados en ese ámbito" (Cortina, 2025, p. 106). En este sentido, la conjunción de los diferentes enfoques éticos en un proceso de desarrollo debe tener como resultado que "los valores de las partes interesadas de una tecnología (de abajo hacia arriba) se sopesan con los valores derivados de marcos teóricos y normativos (de arriba hacia abajo)" (Buijsman, Klenk & van den Hoven, 2025, p. 73). En otras palabras, se trata de identificar los valores éticos deseados para convertirlos en insumo de los procesos de desarrollo y evaluación de los sistemas inteligentes, pero en un sentido constructivo y crítico, no solo instrumental. Esto supone, por ejemplo, una ética de la virtud que permite la incorporación de los marginados de los procesos de construcción de consenso, basados convencionalmente en el estado de las relaciones de poder.

Este aspecto es congruente con un enfoque de coconstrucción de sistemas sociotécnicos. La teoría crítica de la tecnología explica que ésta debe ser analizada a través de los niveles de instrumentalización primaria y secundaria: "El nivel primario simplifica los objetos para su incorporación en un mecanismo, mientras que el secundario integra los objetos simplificados en un entorno natural y social" (Giuliano, 2013, p. 66). No obstante, en este proceso los valores dominantes son agregados al diseño, de manera que se entretejen con los sistemas en forma de "códigos técnicos" y obtienen aceptación social como parte de sus propiedades. Esto significa que el diseño tecnológico no es neutral, y que se pueden encontrar distintas trayectorias de cambio tecnológico de acuerdo con la pugna entre los intereses y valores de diferentes grupos sociales. Precisamente, la idea de la doble instrumentalización de la tecnología abre espacio a la tesis de la racionalización democrática (Feenberg, 2005), que implica que una mayor participación cívica en los procesos de deliberación debe incidir en el desarrollo de tecnologías más responsables.



La combinación de un enfoque crítico de la tecnología y el reconocimiento de la importancia del diseño cívico (Feng & Feenberg, 2008) contribuye a fundamentar las ideas de que es posible incorporar intencionalmente determinados valores éticos en los sistemas, en forma de códigos técnicos y características funcionales de los mismos, de manera que tengan representatividad efectiva los valores y las normas previamente acordadas. Esto contribuiría a fomentar aspectos como "el diseño ético de los procedimientos de toma de decisiones, como una oportunidad para crear sistemas (sociotécnicos) aún más responsables" (Buijsman, Klenk & van den Hoven, 2025, p. 77), y por tanto aplicaciones con mejor desempeño en materia de responsabilidad. Así, parece pertinente concebir el diseño y desarrollo de IA como un proceso de innovación colaborativa (Terrones, 2020), más cercano a las dinámicas de la ciencia y tecnología abiertas (Fecher & Friesike, 2014) que a las propias de la ciencia y tecnología mercantil.

La deliberación pública en torno a problemas de ética y responsabilidad es necesaria para sentar las bases de la selección de valores y criterios de decisión más deseables, pero asimismo es importante la construcción social de dinámicas que permitan su integración en sistemas sociotécnicos que ya encuentran presiones significativas en áreas como políticas de regulación y productividad. En este sentido, resulta pertinente revisar algunos discursos acerca de la ética y la responsabilidad en el desarrollo e implementación de IA, con el fin de adentrarse en el marco del debate sobre este tema.

## 2. Dos discursos sobre la ética de la IA

En el campo del desarrollo tecnológico intervienen muchos factores que resulta necesario ordenar. En este artículo se plantea que la inversión de recursos destinados a la producción de bienes culturales-tecnológicos puede comprenderse desde dos marcos de interpretación: el interés público y el interés privado (y sus correlatos en una ciencia y tecnología de interés cívico y de interés mercantil, respectivamente). Si se reconoce la presencia de ambos marcos de interpretación como conjuntos de valores que orientan los sistemas de acción, entonces se debe admitir que la tensión entre el interés privado y el interés público coloca a la ética en un escenario de conflicto. En tal sentido, se plantea que el desarrollo de IA responsable se encuentra en un dilema: si los valores que promueven la innovación persiguen un fin mercantil, entonces los bienes tecnológicos no pueden tener carácter ético, pero –si se acepta esto– entonces el desarrollo de tecnología no es una actividad que merezca reconocimiento social. En otras palabras, el desarrollo técnico orientado al mercado es un ejercicio instrumental pero no una práctica virtuosa. Debatir acerca de la ética del *software* depende de que se visibilicen los alcances y limitaciones de cada contexto interpretativo.



Un contexto interpretativo es un marco conceptual que le brinda sentido a una situación, y que influye en la manera en que se entienden aspectos como la información y las decisiones (Fuenmayor, 1991). La importancia de los contextos interpretativos reside en que no pretenden ser neutrales, sino que intentan retratar de forma sistemática los valores, intereses y estilos organizacionales que se entremezclan en el campo sociocultural. En ese sentido, también tienen importancia como parte de la lógica de la acción de individuos, grupos e instituciones. Los contextos interpretativos aludidos en este trabajo pueden describirse sucintamente de la siguiente manera:

- Como bien privado, la tecnología es concebida como un activo económico sujeto a las dinámicas de mercado. Esto significa que los derechos de propiedad intelectual (autoría y explotación económica) son exclusivos de sus propietarios, y que los bienes se encuentran sujetos a transacciones comerciales. Los agentes económicos deben maximizar sus beneficios y compiten por dominar el mercado, lo que convierte a la innovación en una

ventaja competitiva. Asimismo, se tiende a minimizar los costos y externalizar las consecuencias negativas de la actividad económica, lo que crea fricciones con entes de regulación y organizaciones civiles. En este contexto, la tecnología se concibe como un bien mercantil.

- Como bien público, la tecnología se considera como un recurso compartido cuyos beneficios deben estar garantizados a toda la sociedad. Los derechos de propiedad intelectual conceden acceso a los bienes cognitivos para favorecer la inclusión. El desarrollo tecnológico está orientado al interés común y se plantea contribuir con el bienestar humano. Asimismo, se fomenta la colaboración y el compartir con miras a facilitar experiencias de apropiación del conocimiento y de innovación. Las políticas tienen como referente la participación ciudadana y la deliberación, lo que garantiza un papel a las instituciones públicas. El concepto de ciencia abierta como bien común recoge algunos aspectos del carácter de la tecnología en este contexto.

En situaciones reales, no se encuentran los casos "puros" que representan los contextos interpretativos. En las relaciones sociales, se despliegan dinámicas de poder que se entretejen con la coconstrucción de sistemas sociotécnicos que implementan tecnologías como la IA desde diferentes aristas (Roca, 2025). Desde una perspectiva convencional, el interés privado tiene el potencial de propiciar la innovación, pero contribuye a crear nuevos problemas sociales; mientras que el interés público favorece la inclusión, pero encuentra dificultades para gestionar eficientemente la inventiva. Los contextos interpretativos que son mutuamente excluyentes permiten delimitar críticamente las maneras en las cuales se practica el desarrollo responsable de IA. La definición de los contextos interpretativos mencionados tiene influencia directa en la selección de los casos de estudio que se ofrecen en este trabajo, en los cuales se intenta exhibir las características de los discursos sobre la ética en el desarrollo y la implementación de IA.



Como se explicó en el apartado anterior, una aproximación convencional a los temas de ética en la IA ha sido enunciar un conjunto de principios y valores que sirvan de orientación a las organizaciones relacionadas con su implementación. En tal sentido, se encuentran numerosas guías institucionales para el reconocimiento de la ética y la responsabilidad de la IA, que coinciden en aspectos como el respeto a los valores de transparencia, justicia y equidad, no maleficencia, responsabilidad y privacidad; pero que también difieren en materia de interpretación e implementación (Buijsman, Klenk & van den Hoven, 2025). En concordancia, el investigador puede proceder críticamente en el sentido de examinar los fundamentos y la pertinencia de los valores y principios enunciados en cada caso.

En el ámbito de las recomendaciones emanadas por entes multilaterales, se tomaron en cuenta las propuestas de la Organización de la Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) y de la Unión Europea (UE). Se puede asignar a tales propuestas un lugar en el contexto interpretativo del conocimiento y la

tecnología como bienes públicos, de manera que su estudio favorece el registro y su comprensión. Asimismo, la lectura de los respectivos documentos permite observar la manera en que sus discursos dialogan con los agentes económicos en la delimitación de los problemas éticos de la IA.

La UNESCO ofrece un marco de referencia para pensar en la ética del desarrollo de IA. Por una parte, comprende la ética como "una reflexión normativa sistemática, basada en un marco integral, global, multicultural y evolutivo de valores, principios y acciones interdependientes" orientada a fomentar "la dignidad humana, el bienestar y la prevención de daños" (2022, p. 9). En cuanto a los sistemas inteligentes, incluye entre las preocupaciones éticas las relacionadas con los ciclos de vida, abarcando aspectos como financiación, investigación, diseño, desarrollo, comercialización, implementación y desmontaje. Asimismo, comprende la IA en un contexto de relaciones entre agentes que contribuyen con su utilización: "los actores de la IA pueden definirse como todo actor que participe en al menos una etapa del ciclo de vida del sistema de IA y pueden ser tanto personas físicas como jurídicas" (2022, p. 9). En tal sentido, la UNESCO comprende los sistemas técnicos como fenómenos arraigados en un contexto social.

La UNESCO (2023) plantea la importancia de fortalecer un enfoque orientado al reconocimiento de los derechos humanos. Los valores y principios desempeñan un papel importante en la recomendación, como orientación para las acciones políticas y jurídicas de los agentes de IA (UNESCO, 2022). Entre los valores que menciona se encuentran: respeto, protección y promoción de los derechos humanos, las libertades fundamentales y la dignidad humana; prosperidad del medioambiente y los ecosistemas; garantía de diversidad e inclusión; vivir en sociedades pacíficas, justas e interconectadas. Asimismo, entre los principios aplicables a la IA se mencionan: 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; gobernanza y colaboración de múltiples partes interesadas.



La propuesta de la Comisión Europea (2019) sobre la implementación de IA confiable también es ampliamente reconocida en diversas fuentes. En esta propuesta, la fiabilidad de la IA se apoya en tres componentes: debe ser lícita (respeto a las leyes), ética (congruente con valores éticos) y robusta (funcional y socialmente estable). Algunos principios éticos que se consideran implicados en el desarrollo e implementación de IA incluyen: respeto a la autonomía humana, prevención del daño, equidad y explicabilidad; atención a las relaciones entre grupos sociales y especialmente a los más vulnerables; y consideración de los posibles riesgos y efectos negativos. A partir de estos aspectos, el documento plantea algunas medidas para avanzar hacia una IA fiable, entre los cuales aparece como principal:

> "Garantizar que el desarrollo, despliegue y utilización de los sistemas de IA cumpla los requisitos para una IA fiable: 1) acción y supervisión

> humanas, 2) solidez técnica y seguridad, 3) gestión de la privacidad y de los datos, 4) transparencia, 5) diversidad, no discriminación y equidad, 6) bienestar ambiental y social, y 7) rendición de cuentas" (Comisión Europea, 2019, p. 2).

En este sentido, la propuesta plantea que se deben impulsar medidas como la investigación abierta y la formación de especialistas, informar sobre las capacidades y las limitaciones de los sistemas inteligentes, facilitar la auditabilidad de los sistemas, trabajar sobre todos los agentes y etapas del ciclo de vida de los sistemas, y resolver las tensiones entre los diferentes principios e intereses (Comisión Europea, 2019). En esta propuesta se da un paso más allá de los principios y valores, y se plantean acciones e indicadores para avanzar en la dirección de entornos de implementación de IA más responsables.

La aproximación de entes multilaterales como la UNESCO y la UE retrata a los agentes de IA como parte de ecosistemas sociotécnicos y no como innovadores aislados, dado que se reconoce la presencia de diversos agentes humanos, intereses y relaciones en contextos reales de implementación. Asimismo, se establecen diferentes valores y principios para orientar las actividades de los agentes que forman parte de los sistemas sociotécnicos, pero especialmente de los productores de la tecnología. En cuanto a su aproximación ética, los entes utilizan una orientación basada en la formulación de normas generales, las cuales se suponen aplicables a todos los casos. No obstante, algunas de las recomendaciones específicas de la UE están más orientadas a las características técnicas de los sistemas, así como al conjunto de relaciones que se entretejen en los contextos de implementación.



En el ámbito de los discursos de los entes privados, se tomaron como referencia las posiciones de Google y Microsoft. Los documentos examinados se incluyen en el contexto interpretativo del conocimiento y la tecnología como bienes privados, de manera que refieren la posición de los agentes económicos frente a las cuestiones éticas que plantean otros organismos. Asimismo, se puede observar cómo aquellos dialogan con otros actores para ofrecer un sentido ético a la implementación de sistemas inteligentes en contextos de mercado.

Google (2024) menciona entre los objetivos del uso de la IA: ser beneficiosa a nivel social; evitar crear o reforzar un sesgo parcial; estar creada y probada para ofrecer seguridad; ser responsable frente a las personas; incorporar principios de diseño de privacidad; mantener estándares altos de excelencia científica; y estar disponible para usos que concuerden con estos principios. Asimismo, se incluye un apartado sobre los usos que no se darán a su IA: tecnologías que puedan causar daño general; armas o cualquier otra tecnología cuyo propósito sea causar o permitir daños a las personas; tecnologías que recopilen o usen información para vigilancia y que infrinjan las normas aceptadas a nivel internacional; tecnologías cuyo propósito desobedezca los principios ampliamente aceptados de leyes internacionales y derechos humanos.

En este caso, la empresa plantea, de forma amplia, los usos que se pueden y no se pueden dar a sus productos de innovación.

Por otra parte, Microsoft (2025) plantea los siguientes principios de desarrollo y uso de IA: equidad, fiabilidad y protección; privacidad y seguridad; inclusión, transparencia y responsabilidad de las personas. En otro documento, la empresa plantea objetivos de desempeño en áreas tales como rendición de cuentas, transparencia, equidad, confiabilidad, privacidad y seguridad, e inclusión (Microsoft, 2022). En ese sentido, la empresa establece un conjunto de principios rectores, pero también incorpora la preocupación por la ética entre los estándares de evaluación de sus productos. Por lo tanto, comunica su interés en convertir los principios en características funcionales de los sistemas.

Los enfoques de empresas como Google y Microsoft identifican la posición pública de las mismas en cuanto que agentes económicos, y responden directamente a preocupaciones éticas sobre el uso de la IA en entornos de implementación. Desde su perspectiva, también se reconoce la IA como parte de un conjunto de relaciones sociales, particularmente en su dimensión institucional. En cuanto a la dirección ética, se puede afirmar que los principios de Google y de Microsoft tienen una expresión orientada a las normas, aunque los objetivos de desempeño del último reflejan un enfoque en las características y los efectos de los sistemas tecnológicos.



En síntesis, en el discurso de los entes multilaterales se incorporan normas que pueden ser relevantes más allá de la implementación de IA, como el respeto a los derechos humanos; o bien, como en el caso de la UE, se hace referencia a principios y medidas de acción para la implementación de sistemas tecnológicos, como la supervisión humana y la rendición de cuentas. El discurso de los entes comerciales también incluye normas generales, más orientadas hacia los sistemas de IA; y como en el caso de Microsoft, se agrega una preocupación por las características y los efectos de los productos que ofrecen. De esta manera, los discursos examinados reflejan los valores e intereses propios de cada organización y, por lo tanto, son coherentes con su perfil y su lugar en el ecosistema digital. En tal sentido, también permiten registrar la manera en que diferentes agentes sociales dan cuenta de su posición ante las preocupaciones que despierta la implementación de sistemas inteligentes en espacios reales, tomando en cuenta la responsabilidad que deben tener en el contexto del debate público.

Es válido cuestionar si todos los principios y valores enunciados son representativos de las preocupaciones éticas de la esfera pública contemporánea. Por ejemplo, algunos movimientos sociales podrían reclamar mayor visibilización de grupos vulnerables en las recomendaciones de la UNESCO (2021), que también se ha pronunciado en favor de una política de ciencia abierta, pero no la menciona en los documentos examinados. Si bien es cierto que los enunciados de una “sociedad justa” e “interconectada” pueden abarcar estos aspectos, no lo hacen explícitamente. En el caso de los entes privados, los valores que proponen son aplicables a todos sus productos. Sin embargo, es conocido que los sistemas tecnológicos no son neutrales, sino que son desarrollados

con intencionalidad y reproducen los sesgos que intervienen en su programación. Así, tomando por caso las críticas recurrentes al manejo de datos de los usuarios por parte de las empresas tecnológicas, debe entenderse que su compromiso con la IA abarca valores distintos a los que incorporan otros productos enmarcados en sus prácticas comerciales. Éstas son solo dos críticas que pueden dirigirse a los discursos basados en principios y valores de la ética en los sistemas de IA.

Un aspecto importante es la manera en que las recomendaciones éticas pueden adaptarse al diseño de *software* como actividad concreta. No parece suficiente que una organización haga declaraciones públicas afirmando su apoyo al desarrollo responsable, sino también debe haber un proceso de adaptación organizativa a los fines de implementar sistemas éticos. Por ejemplo, se pueden crear departamentos sobre la materia, indicadores de calidad de *software*, requerimientos y casos de uso específicamente orientados a cumplir ciertos criterios, consultas a usuarios, etc., medidas cuyos resultados pueden publicarse periódicamente en informes de rendición de cuentas de los agentes involucrados en la IA. Así, es necesario investigar cómo operacionalizar los principios éticos en forma de modelos organizacionales, criterios de decisión y proyectos de desarrollo. Por ejemplo, existen recomendaciones para la implementación responsable de IA en las organizaciones (Wade & Yokoi, 2024), así como propuestas para el desarrollo de IA ética desde el diseño (Gordaliza, 2021).

En síntesis, los discursos acerca de la ética en el desarrollo de sistemas inteligentes parecen reflejar el perfil de las organizaciones que los emiten, y en cierto modo reproducen las lógicas orientadas al bien público y al bien mercantil. Es necesario encontrar discursos compatibles entre sí, que contribuyan a integrar los ecosistemas de desarrollo e implementación a través de iniciativas como, por ejemplo, el diseño participativo (Van der Velden & Mörtberg, 2015), y de ese modo facilitar la incorporación de valores y principios consensuados entre las características funcionales de los sistemas. Esto significa que la ética debe considerarse más que una lista de exigencias o de indicadores de evaluación, y que se debe avanzar hacia la conformación de sistemas sociotécnicos donde la ética y la responsabilidad se consideren relevantes desde las etapas de diseño, de forma que se pueda evaluar su efectividad a lo largo de los ciclos de investigación, desarrollo y aplicación de los sistemas tecnológicos. Precisamente, el enfoque de "ética desde el diseño" (Comisión Europea, 2021; Brey & Dainow, 2023) apunta en la dirección de ofrecer metodologías que permitan evaluar el carácter ético de los sistemas desarrollados a partir de sus etapas iniciales, en lugar de esperar el momento de implementación para observar sus efectos.



### 3. Hacia una ética del diseño de sistemas inteligentes

El desarrollo de IA incorpora la ética desde varias aristas. Aspectos como la limitada eficiencia de la regulación, la opacidad de las empresas en el manejo de los productos informáticos y el impacto de la implementación de IA generan preocupación en la esfera pública (AI Index Steering Committee, 2024). La ética de las tecnologías surge

como una preocupación compartida, en torno a la cual se pueden tejer acuerdos que favorezcan a todas las partes interesadas. Sin embargo, no hay más que mirar el devenir del desarrollo tecnológico para caer en cuenta de la prevalencia que tienen las prácticas de mercado en la adopción social de tecnologías digitales. El caso de la IA generativa resulta elocuente en este sentido: la puja entre diferentes desarrolladores de modelos de lenguaje retrata la competencia que existe entre las empresas por convertirse en los proveedores de facto de una tecnología que está cambiando la manera en que se utilizan las interfaces digitales. Esta constante en los procesos de incorporación de innovaciones tecnológicas, sin la revisión crítica de todas las variables, ha hecho que se hayan dado lugar a daños irreversibles a terceros, como en el caso de la degradación ambiental.

En lo que respecta a la introducción de valores éticos en el desarrollo tecnológico, se ha considerado la importancia de "descubrir orientaciones éticas interculturales para utilizar con bien los medios tecnocientíficos, pero también diseñar guías éticas para incorporarlas en los sistemas inteligentes de forma que actúen éticamente" (Cortina, 2024, p. 29). En ese sentido, puede plantearse que las lecciones en materia de normas universales y beneficios de los bienes tecnológicos contribuyan a fomentar una suerte de "práctica virtuosa" del desarrollo tecnológico, donde valores como la justicia y la responsabilidad sean reconocidos como parte de las virtudes de los desarrolladores y de los bienes internos de los productos de innovación. Así, podría transferirse el reconocimiento de valores compartidos a métodos de desarrollo que contribuyan a recrear un modo no instrumentalista de gestión tecnológica, con miras a transformar los procesos, las relaciones y los resultados de investigación. De esa manera, resulta necesario plantear cuáles son los valores, principios y términos de actuación que permitirán confirmar que un determinado producto es el resultado de un conjunto de acuerdos de contenido ético.



En términos de bienes públicos y bienes privados (ciencia abierta y ciencia mercantil), esta situación dibuja un escenario de conflicto, donde se hace patente la ausencia de marcos de interpretación de la tecnología coherentes con las necesidades de sociedades pluralistas. Por una parte, se puede recurrir a la tradición ética para encontrar discursos en torno a las virtudes de los agentes tecnológicos, las normas universales o las consecuencias de la acción ética. Pero, asimismo, los discursos gubernamentales y los intereses de mercado se combinan y se contrarrestan en el crisol de las relaciones políticas y económicas. Además, las prácticas de negocio de las empresas logran convencer a los usuarios, quienes, por medio del consentimiento explícito, parecen aceptar los términos de uso de las aplicaciones a cambio de costos percibidos como males menores. El resultado es un escenario complejo, en el cual las tecnologías digitales se han convertido en dispositivos no neutrales con capacidad para impactar en las relaciones sociales y culturales sin vocación ética ni responsabilidad (Suárez, 2024).

En este contexto, puede criticarse que los problemas que generan el desarrollo e implementación de IA parecen menos dilemas éticos que contradicciones de la

planificación del mercado. En un escenario de negociación, la ética puede representar la posición de entes reguladores y agentes de mercado con miras a definir acuerdos que permitan articular la intervención de las tecnologías digitales en los procesos sociales, como suele ocurrir con el ascenso de nuevos paradigmas organizacionales (Pérez, 2020). En ese sentido, también constituye una alternativa para participar en la definición de políticas públicas y pautas de gobernanza de los ecosistemas digitales. Los entes multilaterales y las empresas que se tomaron como referencia hacen su parte al expresar posiciones acordes con los derechos fundamentales o con la recepción de sus bienes de mercado, respectivamente. En contraste, las diferencias pueden comenzar a aparecer en los niveles de gobernanza de los ecosistemas digitales y en las prácticas de diseño de las aplicaciones. Es decir, en los ecosistemas sociotécnicos de la digitalidad.

Tanto la UNESCO como la UE siguen una orientación normativa que se propone guiar políticas y acciones, basada en temas como los derechos humanos y la sostenibilidad. En consecuencia, se enfocan en la definición de principios éticos universales como equidad, privacidad, responsabilidad y no discriminación en el desarrollo de IA, y se encuentran en sintonía con marcos de principios establecidos previamente (Floridi, 2023). La UNESCO tiene una perspectiva global y multicultural, centrada en los derechos humanos, la diversidad y la sostenibilidad, mientras que la UE ofrece un enfoque regional que prioriza aspectos técnicos como la legalidad, ética y robustez de los sistemas técnicos, incluyendo sus condiciones de fiabilidad. Por otra parte, Google y Microsoft también comparten principios éticos como equidad, privacidad, transparencia y responsabilidad en el diseño de IA, así como su preocupación por la no maleficencia y la seguridad de los sistemas. Google prioriza ciertos beneficios sociales, como la excelencia científica, a la vez que restringe algunos usos de la IA que pueden perjudicar al ser humano, mientras que Microsoft enfatiza aspectos técnicos como la seguridad e incorpora estándares de medición de desempeño.



En este sentido, todos los agentes de IA comparten principios éticos como transparencia, equidad, privacidad y responsabilidad, además de que reconocen el impacto social de las tecnologías a través de riesgos como la discriminación y la seguridad. La UNESCO y la UE son entidades cuya razón de ser las obliga a preocuparse por el bien público, mientras que Google y Microsoft son entidades privadas que deben encontrar un balance entre sus opciones éticas, iniciativas de innovación y oportunidades de mercado. Por lo tanto, las propuestas de la UNESCO y la UE pretenden ser aplicables en diferentes contextos, mientras que las de Google y Microsoft están orientadas a la recepción de sus prácticas comerciales. En perspectiva, los marcos de la UNESCO y la UE adoptan una ética deontológica (basada en la idea de bienes universales), mientras que los marcos de Google y Microsoft combinan deontología con consecuencialismo (se proponen maximizar los beneficios y reducir los perjuicios). El problema de la virtud de los agentes tecnológicos queda implícito, y su lugar lo ocupan los atributos de los bienes tecnológicos que deberían desarrollarse con criterios éticos. En otras palabras, se presume que un desarrollador responsable es aquel que produce sistemas de IA bajo criterios de responsabilidad.

De este modo, ambos enfoques pueden considerarse tanto conflictivos como complementarios. Contemplados como parte de un cuadro más amplio, las diferentes perspectivas de ética en el desarrollo de IA abarcan aspectos comunes. Se parte de que la IA debe contribuir con el bienestar humano, evitar daños y fomentar valores compartidos. Además, se reconocen temas clave como el respeto a la dignidad humana; la equidad y la no discriminación; la transparencia, la responsabilidad y la sostenibilidad. Estos aspectos son representativos de una ética mínima, que puede traducirse en políticas, marcos jurídicos y guías para el diseño de productos técnicos en entornos complejos. En concordancia, se estaría integrando los aportes de los niveles normativos (global), regulatorios (regional) y operativos (privado) en la definición de bienes cognitivos, avanzando hacia la determinación de su concepción ética, su alcance jurídico, sus características técnicas y sus restricciones de adopción en el mercado. Sin embargo, en este escenario subyace una suerte de sesgo de selección: los diferentes discursos no son complementarios *per se*, sino como consecuencia de la manera en que fueron caracterizados para este estudio. En realidad, se trata de discursos en conflicto que solo pueden reconciliarse a través de iniciativas de gobernanza global.

Otro escenario puede basarse en lo que cabe esperar de la innovación como proceso de interés público. La ciencia y tecnología abiertas ofrecen un conjunto de técnicas de gestión de los bienes cognitivos que permiten que sean administrados colectivamente de acuerdo con ciertas normas (Hess & Ostrom, 2016). Por ejemplo, un repositorio de publicaciones de acceso abierto puede ser manejado por una fundación, y los artículos almacenados cuentan con licencias que permiten su reutilización. En el caso de la IA, se ha recomendado tomar en cuenta una definición de código abierto que incorpore los valores de "autonomía, transparencia, reutilización sin fricciones y mejora colaborativa" en el desarrollo de modelos inteligentes (Open Source Initiative, 2025), tal como se establece para otros productos de *software*. De este modo, los modelos de lenguaje, por ejemplo, desarrollados a través de la colaboración entre diferentes actores y entrenados con bases de datos de dominio público, pueden ofrecerse con licencias abiertas como una manera de retribuir a la sociedad. Así, la ciencia abierta resulta ser más que una práctica que persigue la excelencia de los bienes internos, y representa una contribución a la transformación de marcos institucionales, procesos técnicos y productos de conocimiento, ciencia y tecnología, en el mismo sentido en que una ética aplicada a la IA tiene su anclaje en un conjunto de principios éticos generales.



Se deduce entonces que, para superar la dicotomía entre bienes públicos y privados, resulta necesario democratizar los espacios de gestión de los bienes tecnológicos, considerando, entre otros aspectos, que sería una alternativa para someter a escrutinio los intereses, valores y principios éticos implícitos en los procesos de diseño, desarrollo, implementación y apropiación del conocimiento. En otras palabras, los aportes normativos, regulatorios y operativos en la definición de bienes cognitivos deben tener una base cívica, que puede ser cultivada en espacios de ciencia abierta y ciencia ciudadana, pero que también cuenta con todos los

mecanismos de formación de la opinión pública. La democratización de la gestión de bienes tecnológicos representaría una alternativa para el establecimiento de criterios de ética y responsabilidad en su diseño y liberación, no solo por gracia de aspectos como la auditabilidad de las iniciativas privadas y públicas, sino también por su efecto en la consolidación de una ética del diseño participativo que contribuya a modelar las prácticas de desarrollo de sistemas informáticos. En tal sentido, la racionalización democrática puede dar lugar no tanto a la implementación de otros sistemas técnicos, sino sobre todo a la institucionalización de otros sistemas éticos.

Superar la prevalencia de los discursos público y privado, y propiciar la adopción de un discurso comunitario basado en la ciencia abierta, requiere mantener una dialéctica entre la reflexión sobre los deberes (deontología) y los efectos de la adopción de los sistemas tecnológicos (consecuencialismo), pero también exige la incorporación de virtudes éticas en el carácter de los agentes tecnológicos. Por lo tanto, la deliberación acerca de la ética de los agentes y las narrativas que la sostienen no puede postergarse indefinidamente. En la medida en que los sistemas sociotécnicos son complejos, la posición ética debe estar comprometida a responder por aspectos como los conflictos de intereses, los efectos inesperados de las tecnologías y la realidad de los diferentes contextos de aplicación. Entre los agentes de IA, para utilizar la terminología de la UNESCO, se deben incluir participantes ocupados de las normas, las regulaciones, las prácticas de negocio y el deber cívico en contextos regionales y globales, con los fines de que la crítica a las relaciones mediadas por la tecnología sirva de fundamento para la aprobación de iniciativas acordes con valores consensuados como el respeto a la dignidad humana, así como con el diseño de modelos organizacionales y métodos de desarrollo congruentes con los mismos.



En tal sentido, las herramientas conceptuales y metodológicas de la ciencia y tecnología abiertas pueden contribuir a orientar el manejo de bienes cognitivos mercantiles en la dirección de los bienes comunes. Por ejemplo, puede reivindicarse el carácter ético de las prácticas de desarrollo de IA (ética de la virtud) y crear acuerdos que permitan la incorporación de valores responsables en procesos de diseño tecnológico orientados a los usuarios, definidos a su vez como sujetos éticos (racionalización democrática). Asimismo, se pueden crear prácticas de participación cívica que ofrezcan espacio a la democratización de los procesos de implementación de tecnologías, salvaguardando la perspectiva ética frente a la aceptación del mal menor o las imposiciones de la autoridad. En tal sentido, el desarrollo ético y responsable de sistemas inteligentes pasa necesariamente por la cocreación de sistemas sociotécnicos donde tengan cabida la ética y la responsabilidad como parte del diseño y la implementación de tecnologías.

## Conclusiones

Es claro que las falencias éticas de la IA no son imputables a la propia IA, al menos como producto terminado. El diseño y desarrollo de programas informáticos es un

proceso reflexivo y orientado subjetivamente donde toman parte diferentes intereses. La arquitectura y los términos de uso de las aplicaciones responden a esquemas de diseño que favorecen el modelo de negocio de las empresas. Así, pueden introducir especificaciones funcionales que no responden solamente a necesidades de eficacia y eficiencia. Además, una vez que los usuarios comienzan a utilizar una aplicación, los desarrolladores obtienen más información sobre sus gustos y pueden realizar cambios en el programa. En el caso de la IA generativa, se agrega el hecho de que los sistemas inteligentes son entrenados con textos abiertos, por lo que tienden a incorporar sesgos humanos, como por ejemplo con respecto a prejuicios sociales. La cuestión de los valores de los desarrolladores de IA confronta solo una parte del problema. Como ecosistemas sociotécnicos, los entornos de innovación deben fundamentarse en principios éticos que contribuyan con la deliberación sobre los verdaderos intereses y valores implícitos en el desarrollo de bienes tecnológicos.

Como señala una perspectiva crítica, los valores e intereses implícitos en las aplicaciones digitales tienden a normalizarse en contextos de uso. Los métodos de diseño convencionales se combinan con requerimientos de desarrollo derivados de los intereses de los agentes de mercado en el marco de la lógica propia del contexto interpretativo de la tecnología mercantil. Sin embargo, los mismos procesos de diseño, desarrollo e implementación pueden utilizarse para liberar programas donde tengan cabida valores éticos como la responsabilidad, intereses cívicos como la representación de grupos vulnerables y adaptaciones que respondan al bien común antes que al monopolio del mercado. La racionalización democrática de la tecnología digital es posible si los intereses cívicos ganan lugar en los espacios de debate sobre el diseño tecnológico, y esto se traduce en la adopción de esquemas de diseño participativo. Por ejemplo, prácticas de la ciencia y tecnología abierta, como los laboratorios ciudadanos, la documentación abierta y el *software* de código abierto pueden convertirse en mecanismos para reunir intereses comunitarios alrededor del diseño de *software* en entornos éticos. Pero, asimismo, la democratización implica otros aspectos como el escrutinio público de las iniciativas de innovación y la participación cívica en los procesos de regulación pública.



La deliberación abierta sobre los principios éticos que deben fundamentar las actividades de desarrollo tendría el papel de favorecer el consenso alrededor de los valores, intereses, prácticas y relaciones acordes con los mismos. La ética, distinta de la opinión general y de las decisiones de la autoridad, contribuye a establecer criterios de evaluación de la acción y ayuda a fundamentar acuerdos sobre los cuales pueden hacerse reclamos de responsabilidad. La introducción de la ética en el diseño debe responder qué valores y principios serán transferidos en forma de características funcionales a las aplicaciones tecnológicas, sean sistemas inteligentes o agentes autónomos. Pero, asimismo, implica un marco de conducta para los agentes de la IA, quienes tienen oportunidad de comprometerse en la coconstrucción de sistemas sociotécnicos más éticos y responsables. En ese sentido, el diseño ético apunta a fomentar un conjunto de prácticas virtuosas ocupadas de los bienes internos y la excelencia en el desarrollo de aplicaciones, pero puede contribuir a reforzar sistemas

sociotécnicos donde tengan lugar relaciones de responsabilidad mutua y apoyar un conjunto de valores que tengan arraigo más allá de los contextos restringidos de la ética aplicada, con el fin de propiciar otros modos de comprender la tecnología.

Es claro que este horizonte no implica una trayectoria lineal, sino que se requiere del diálogo entre todas las partes. Los valores éticos que pudieran proponer los académicos deben contrastarse con los valores culturales en cada contexto social. Las diferencias entre agentes que responden a la lógica del bien público y a la lógica mercantil pueden revertirse desde la base de la complementariedad que cabe en los espacios participativos de la ciencia abierta. Existen numerosas carencias desde el punto de vista material, como las dificultades para reunir capacidades de financiamiento y escalamiento para la innovación fuera de los contextos de mercado. Sin embargo, al delimitar una situación ideal es posible contribuir con la conformación de un marco de diseño que facilite que el problema de la ética en la IA sea una cuestión de base, y no solo un pretexto para establecer marcos de negociación con las empresas una vez que se perciben los efectos indeseados de la tecnología en la sociedad. El dilema de la ética en la IA puede resolverse *ab initio* desde la perspectiva de un diseño responsable y abierto a la participación cívica.

**Bibliografía**



Adams, R., Adeleke, F., Florido, A., de Magalhães Santos, L. G., Grossman, N., Junck, L. & Stone, K. (2024). Global Index on Responsible AI 2024. Recuperado de: https://agatadata.com/wp-content/uploads/2024/11/Global-Index-on-Responsible-AI-2024-Corrected-Edition-25_10_24-Gobierno-de-Canada-y-USAID.pdf.

AI Index Steering Committee (2024). The AI Index 2024 Annual Report. Standford: Stanford University. Recuperado de: https://aiindex.stanford.edu/report/.

Askell, A., Brundage, M. & Hadfield, G. (2019). The Role of Cooperation in Responsible AI Development. Recuperado de: https://arxiv.org/pdf/1907.04534.

Brey, P. & Dainow, B. (2024). Ethics by design for artificial intelligence. AI Ethics, 4, 1265–1277. DOI: https://doi.org/10.1007/s43681-023-00330-4.

Buijsman, S., Klenk, M. & van den Hoven, J. (2025). Ethics of AI. Toward a “Design for Values” Approach. En N. Smuha (Ed.), The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence (59-78). Cambridge: Cambridge University Press.

Canca, C. (2020). Operationalizing AI Ethics Principles. Communications of the ACM, 63(12). DOI: https://doi.org/10.1145/3430368.

Coeckelbergh, M. (2021). Ética de la inteligencia artificial. Madrid: Cátedra.

Comisión Europea (2019). Directrices éticas para una IA fiable. Recuperado de: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Cortina, A. (2024). ¿Ética o ideología de la inteligencia artificial? El eclipse de la razón comunicativa en una sociedad tecnologizada. Madrid: Paidós.

European Commission (2021). Ethics By Design and Ethics of Use Approaches for Artificial Intelligence. Recuperado de: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf.

Fecher, B. & Friesike, S. (2014). Open Science: One Term, Five Schools of Thought. Opening Science. The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing. Springer.

Feenberg, A. (2005). Teoría crítica de la tecnología. Revista Iberoamericana de Ciencia, Tecnología y Sociedad, 2(5), 109-123. Recuperado de: https://ojs.revistacts.net/index.php/CTS/article/view/1018.

Feng, P. & Feenberg, A. (2008). Thinking about Design: Critical Theory of Technology and the Design Process. En P. Kroes, P. Vermaas, A. Light, S. Moore (Eds.), Philosophy and Design (105-118). Dordrecht: Springer. DOI: https://doi.org/10.1007/978-1-4020-6591-0_8.



Floridi, L. (2023). The Ethics of Artificial Intelligence. Principles, Challenges, and Opportunities. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/oso/9780198883098.001.0001.

Fuenmayor, R. (1991). Truth and openness: An epistemology for interpretive systemology. Systems Practice, 4(5). Recuperado de: http://www.saber.ula.ve/bitstream/handle/123456789/15918/fuenmayor-truth.pdf.

Giuliano, H. (2013). La teoría crítica de la tecnología: una aproximación desde la ingeniería. Revista Iberoamericana de Ciencia, Tecnología y Sociedad, 8(24), 63-74. Recuperado de: https://ojs.revistacts.net/index.php/CTS/article/view/626.

Google (2024). Principios de la IA. Objetivos para crear IA beneficiosa. Recuperado de: https://ai.google/static/documents/ES-419-AI-Principles.pdf.

Gordaliza, E. (2021). Inteligencia Artificial Responsable desde el diseño. Recuperado de: https://www.bbva.com/wp-content/uploads/2021/06/InteligenciaArtificialResponsable_EstherDelaTorre_BBVA.pdf.

Habermas, J. (1999). Teoría de la acción comunicativa. I. Racionalización de la acción y racionalidad social. Madrid: Trotta.

Hess, C. & Ostrom, E. (2016). Los bienes comunes del conocimiento. Madrid: Traficantes de Sueños.

Lauwaert, L. & Oimann, A. (2025). Moral Responsibility and Autonomous Technologies. Does AI Face a Responsibility Gap? En N. Smuha (Ed.), The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence (101-116). Cambridge: Cambridge University Press.

MacIntyre, A. (2013). Tras la virtud. Madrid: Crítica.

Microsoft (2025). Promoción de prácticas de inteligencia artificial responsables. Recuperado de: https://www.microsoft.com/es-es/ai/responsible-ai.

Microsoft (2022). Microsoft Responsible AI Standard. Recuperado de: https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf.

Misselhorn, C. (2022). Artificial Moral Agents: Conceptual Issues and Ethical Controversy. En S. Voeneky, P. Kellmeyer, O. Mueller & W. Burgard (Eds.), The Cambridge Handbook of Responsible Artificial Intelligence. Interdisciplinary Perspectives (31-49). Cambridge: Cambridge University Press.

OpenAI (S.F.) OpenAI Charter. Recuperado de: https://openai.com/charter/.



Open Source Initiative (2025). The Open Source AI Definition – 1.0. Recuperado de: https://opensource.org/ai/open-source-ai-definition.

Pérez, C. (2020). Revoluciones tecnológicas y paradigmas tecnoeconómicos. En D. Suárez, A. Erbes & M. F. Barletta (Eds.), Teoría de la innovación, evolución, tendencias y desafíos: herramientas conceptuales para la enseñanza y el aprendizaje (134-159). Madrid: Ediciones Complutense & UNGS.

Roca, S. (2025). Defectuoso por diseño: Relaciones de poder en la esfera digital. Teknokultura. Revista de Cultura Digital y Movimientos Sociales, 22(1), 69-76. DOI: https://doi.org/10.5209/tekn.95708.

Suárez, J. L. (2023). La condición digital. Madrid: Trotta.

Terrones, A. (2020). Inteligencia artificial, responsabilidad y compromiso cívico y democrático. Revista Iberoamericana de Ciencia, Tecnología y Sociedad, 44(15), 253-276. Recuperado de: https://ojs.revistacts.net/index.php/CTS/article/view/166.

UNESCO (2021). Proyecto de recomendación sobre la ciencia abierta. Recuperado de: https://unesdoc.unesco.org/ark:/48223/pf0000378841_spa.

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

UNESCO (2023). Key facts UNESCO's Recommendation on the Ethics of Artificial Intelligence. Recuperado de: https://unesdoc.unesco.org/ark:/48223/pf0000385082.

Van den Hoven, J., Vermaas, P. & Van de Poel, I. (Ed.) (2015). Handbook of Ethics, Values, and Technological Design. Sources, Theory, Values and Application Domains. Reino Unido: Springer. DOI: https://doi.org/10.1007/978-94-007-6970-0.

Van der Velden, M. & Mörtberg, Ch. (2015). Participatory Design and Design for Values. En J. van den Hoven, P. E. Vermaas e I. van de Poel (Eds.), Handbook of Ethics, Values, and Technological Design. Sources, Theory, Values and Application Domains (41-66). Dordrecht: Springer.

Wade, M. & Yokoi, T. (2024). How to Implement AI — Responsibly. Harvard Business Review, 10 de mayo. Recuperado de: https://hbr.org/2024/05/how-to-implement-ai-responsibly.

# Los “problemas de la inteligencia artificial”. Desafíos éticos y políticos en la era digital

# Os “problemas da inteligência artificial”. Desafios éticos e políticos na era digital

# *The “Problems of Artificial Intelligence”. Ethical and Political Challenges in the Digital Age*

**María del Mar Monti** *

Este artículo indaga en los “problemas de la IA”, cómo se definen y a través de qué discursos se despliegan, colocando el foco en las relaciones asimétricas de poder que moldean su diseño, desarrollo y usos. En particular, aborda tres problematizaciones que son parte de las agendas gubernamentales y de la sociedad: i) el atravesamiento de los sistemas de IA en la vida cotidiana de las personas y sus formas de adopción; ii) la convergencia tecnológica y los debates sobre la recolección y utilización de los datos personales; y iii) la creciente influencia de los discursos hegemónicos de las grandes corporaciones tecnológicas en la configuración de la esfera pública digital. Finalmente, se reflexiona sobre la urgencia de proteger y garantizar el pleno ejercicio del derecho humano a la ciencia en cuanto a posibilidades de acceso a las tecnologías digitales y, principalmente, a la construcción de instancias de participación para la toma de decisiones sobre su diseño, desarrollo, uso y regulación.

**Palabras clave:** inteligencia artificial; ética de la IA; gobernanza tecnológica; derechos humanos; asimetrías de poder

Este artigo tem como objetivo analisar os “problemas da IA” investigando como são definidos e por meio de quais discursos se manifestam, com ênfase nas relações de poder assimétricas que moldam seu desenho, desenvolvimento e uso. Em particular, examina três eixos de problematização que passaram a integrar as agendas governamentais e sociais: i) a inserção dos sistemas de IA na vida cotidiana das pessoas e as formas de sua adoção; ii) a convergência tecnológica e os debates em

* Magíster en estudios sociales de la ciencia y la tecnología, Universidad de Salamanca, España. Investigadora del Grupo Política y Gestión, Facultad de Ciencia Política y Relaciones Internacionales, Universidad Nacional de Rosario, Argentina. Cofundadora y analista de tecnologías emergentes, CoRegener8 GmbH, Suiza. Contacto: mdmmonti@coregener8.com. ORCID: https://orcid.org/0000-0001-6672-6267.

torno da coleta e do uso de dados pessoais; e iii) a crescente influência das grandes corporações tecnológicas na configuração da esfera pública digital e na consolidação de discursos hegemônicos. Por fim, o artigo reflete sobre a urgência de proteger e assegurar o pleno exercício do direito humano à ciência, tanto no que se refere ao acesso às tecnologias digitais quanto, sobretudo, à construção de instâncias de participação para a tomada de decisões relativas ao seu desenho, desenvolvimento, uso e regulamentação.

**Palavras-chave:** inteligência artificial; ética da IA; governança tecnológica; direitos humanos; assimetrias de poder

*This article delves into the "problems of AI", focusing on how they are defined and articulated through discourse. It analyses the asymmetrical power relations that shape their design, development, and application. The discussion highlights three key issues currently central to governmental and social agendas: i) the integration of AI into everyday life and patterns of adoption; ii) technological convergence and the implications of personal data collection and use; and iii) the role of big tech in structuring the digital public sphere and hegemonic discourse. This article concludes with an urgent call to safeguard the human right to science by ensuring equitable access to digital technologies and fostering participation in decision-making processes concerning their design, development, use, and regulation.*



***Keywords:*** *artificial intelligence; AI ethics; technology governance; human rights; power asymmetries*

## Introducción

Este artículo aborda la problematización de las tecnologías digitales, en particular de los sistemas de inteligencia artificial (IA), en las agendas gubernamentales y sociales. Se parte de considerar que el análisis de estas tecnologías es indisociable del contexto en que surgen, se desarrollan y utilizan, y de los debates éticos que se generan en torno a ellas. Es decir, mientras responden a las condiciones técnicas de su tiempo, están atravesadas por los recursos de poder, valores y discursos dominantes de los actores involucrados en las diferentes etapas del ciclo de vida de cada tecnología.

Siguiendo a Pedace *et al.* (2023, p. 3), se conceptualiza a la IA como modelos algorítmicos y sistemas capaces de procesar grandes volúmenes de información y de mejorar su desempeño a partir de la experiencia, superando en muchos casos las funciones para las cuales fueron originalmente programados. Asimismo, se recupera elementos de la *Recomendación sobre ética de la IA* de la UNESCO (2022), especialmente la idea de que estos sistemas operan con diferentes grados de autonomía y pueden generar resultados que influyen en la toma de decisiones en entornos tanto materiales como virtuales. De este modo, la ética se constituye como "una base dinámica para la evaluación y orientación normativa de la IA, tomando como referencia la dignidad humana, el bienestar y la prevención de daños, en consonancia con la ética de la ciencia y la tecnología" (UNESCO, 2022, p. 10).



En un escenario caracterizado por la presencia de sistemas de IA en casi todos los aspectos de la vida cotidiana, es fundamental abordar sus implicancias desde un enfoque crítico y multidimensional que trascienda las aproximaciones meramente técnicas. El objetivo del artículo es indagar en los "problemas de la IA", cómo se definen y a través de qué discursos se despliegan, para luego avanzar en el análisis de situaciones problemáticas actuales que representan diversos desafíos éticos y políticos para los gobiernos y las sociedades. En esta labor, se recurrió a herramientas del análisis de políticas públicas y a una revisión bibliográfica especializada, procurando visibilizar la producción académica en idioma español e incorporar de manera equitativa la perspectiva de autoras.

Tal como se ha señalado, no se pretende aquí realizar una exploración exhaustiva de las intersecciones entre IA y ética, sino que la atención se coloca en tres problematizaciones. En primer lugar, la adopción cotidiana y en gran medida acrítica de sistemas de IA, lo que Sadin (2024) define como el acompañamiento algorítmico de la vida. En segundo lugar, la convergencia tecnológica –entre los sistemas de IA y las neurotecnologías–, y las asimetrías de poder en la recolección y utilización de los datos personales. Finalmente, la creciente influencia de los discursos hegemónicos de las grandes corporaciones tecnológicas en la configuración de la esfera pública digital, y el consecuente debilitamiento de consensos políticos y morales básicos. Ante este panorama, se plantea la urgencia de proteger y garantizar el pleno ejercicio del derecho humano a la ciencia, tanto en términos de acceso como de participación

democrática en la toma de decisiones sobre el diseño, el desarrollo, el uso y la regulación de estas tecnologías.

## 1. La IA como problema y como solución

Existe un amplio consenso en la comunidad académica respecto de la naturaleza transformadora de las tecnologías digitales, que han inaugurado un nuevo capítulo en la historia de la humanidad. En palabras de Innerarity: "La transformación digital es una de las fuerzas más poderosas que están moldeando el futuro de nuestras sociedades" (2025, p. 6). Este disparador puede ser leído desde múltiples dimensiones interrelacionadas: i) la dimensión técnica, vinculada al diseño, el desarrollo y la adopción de tecnologías emergentes como la IA; ii) la dimensión económica con el advenimiento del capitalismo de plataformas y la progresiva concentración del poder en las grandes empresas tecnológicas o *big tech*;[1] iii) la dimensión social, que recupera las preocupaciones centrales de la filosofía de la tecnología: los vínculos entre sociedad y tecnologías, los contextos donde éstas se producen y difunden, y sus impactos; y iv) la dimensión geopolítica, ante la redefinición de los centros de poder en el ámbito digital, la "carrera tecnológica", los grises en la gobernanza global de estas tecnologías y sus efectos en la soberanía digital de los Estados nacionales.



En este artículo, la digitalización es entendida no solo como un fenómeno tecnológico, sino principalmente como un proceso social y político. Cada una de estas dimensiones remite a discusiones que en la actualidad son parte de diferentes agendas. No será objetivo de este trabajo puntualizar en cada una de ellas, sino mostrar la complejidad del contexto y la incertidumbre frente a un panorama donde la velocidad parece imponerse a la reflexividad. Estas cuestiones están siendo problematizadas por diversos actores –sociales, políticos, económicos–, con diferentes recursos de poder y urgencias. Sin embargo, comparten un elemento aglutinador: todas están atravesadas por desafíos éticos.

En la época donde se intensifican las verdades indiscutibles, el propio concepto de ética es utilizado, en ocasiones, para oficiar de contrafuego de las peores derivas. En las antípodas de la definición de ética de Spinoza, "en cuanto la realización de sí: sin un objetivo predador y en armonía con el resto y con el medio" (Sadin, 2024, p. 24). Se evidencia la necesidad de promover un análisis crítico acerca de la incorporación de la ética en los debates sobre tecnologías emergentes, particularmente en el ámbito de la IA. Estos debates suelen centrarse en los denominados "problemas de la IA", cuya definición, en muchos casos, resulta difusa.

[1] La capacidad de control de la IA por parte de unas pocas firmas no tiene precedentes. En particular, se hace referencia a Amazon, Google, Meta y Microsoft, aunque Alibaba, Tencent y otros gigantes chinos, como la empresa dueña de TikTok, también desarrollan y utilizan IA apropiándose de código y datos a mansalva (Rikap, 2023, p. 72).

La definición de los problemas no es poco significativa ni mucho menos neutral. Cuando se nombra un problema, además de ponerse de manifiesto cuáles son los aspectos que se juzgan insatisfactorios, aparece prefigurada cuál es la salida deseable o la perspectiva desde la cual interesa transformar esa realidad. En otras palabras, los problemas no están dados; no existen solo objetivamente en tanto tales, sino que requieren de la percepción de un actor individual o colectivo que se plantee la cuestión y la problematice socialmente. En la construcción del problema entran en juego valores que expresan y definen los ámbitos de pertenencia de los actores, donde éstos se forman e informan (Díaz, 1997). Por consiguiente, la capacidad de los actores afectados por una misma problemática para movilizar recursos e influir en su incorporación en la agenda es, a menudo, desigual.

Siguiendo a Diaz (1997, p. 9), la "construcción de la agenda supone la emergencia del problema, su definición y su inserción en el conjunto de cuestiones priorizadas en el programa de decisión y actuación del poder público", lo que se define como agenda gubernamental. Sin embargo, es preciso reconocer que esta agenda convive con otra, la agenda social o sistémica, que incluye los asuntos que los miembros de una sociedad consideran relevantes, merecedores de la atención pública. Ambas agendas están conectadas entre sí, "hay una agenda de los ciudadanos […] que puede preceder y determinar la agenda del gobierno o ser inducida por las preocupaciones y prioridades gubernamentales, que puede empatar con la del gobierno o diferir de ella en mayor o menor grado" (Aguilar Villanueva, 1993, p. 31). El análisis de las agendas permite reconstruir, a partir de las temáticas priorizadas, la escala de valores y preferencias sociales, las imposiciones políticas, los cálculos racionales de conveniencia, los intentos de anticipar o limitar la definición por parte de otros actores, así como las urgencias derivadas de instancias ajenas.



En el transcurso de los últimos años, se ha observado una creciente incorporación de la IA como "cuestión socialmente problematizada" (Oszlak & O'Donnell, 1976) en las agendas de gobierno –nacionales e intergubernamentales– y de la ciudadanía. La manera en que emergió como problema es diversa y se vincula con los valores, necesidades y urgencias de los actores que lo movilizan. Es preciso señalar en este punto que quienes promueven la cuestión –o se ven afectados por ella– pueden no ser los mismos en cada momento del ciclo de vida de la tecnología.

Queda claro, entonces, que la manera en que se define el problema entraña su posible solución. En palabras de Majone (1997), cada actor construirá sus argumentos en función de sus respectivos intereses y de la lectura que hace de su entorno, interviniendo en la marcha de los asuntos colectivos con el fin de mantener, reformar o transformar el orden social y político. Los CEO de las *big tech* han jugado un rol clave en la problematización social de la IA. Rikap (2023) señala que estos líderes definen a la IA como una tecnología que sustituye e incluso supera a la inteligencia humana, dejando en segundo plano la posibilidad de entenderla como complementaria. Inicialmente, muchas de estas empresas adoptaron discursos que reconocían los posibles riesgos éticos. Sin embargo, más allá de los matices, estas corporaciones sostienen que la

regulación debe focalizarse en la fase de adopción, ya que el problema no se encuentra en el diseño o desarrollo de la tecnología sino en cómo se usa. De este modo, desvían la atención del proceso de producción que permanece bajo su control exclusivo. En otras palabras, "un excesivo foco en la agencia de la IA corre el riesgo de pasar por alto el papel de los agentes que controlan la IA" (Rikap, 2023, p. 81).

La definición de la cuestión en términos de progreso lineal, con énfasis en los beneficios derivados de la innovación y en la competitividad, se cimenta en la intención de consolidar sus liderazgos y posicionamientos por encima de los debates éticos. Ello puede explicar, en parte, la falta de apoyo de las *big tech* al llamado a pausar el desarrollo de sistemas más avanzados que GPT-4, promovido en la carta "*Pause Giant AI Experiments"* (The Future of Life Institute, 2023). Este pedido de pausa por seis meses se fundamentaba en la cuestión del "potencial riesgo para la humanidad". En la Conferencia de Asilomar (2017), organizada por el mismo instituto, ya se había discutido la necesidad de alinear los valores y comportamientos de los sistemas de IA con los valores humanos. Esta conferencia concluyó con la formulación de 23 principios que suponen, para algunos, un valioso punto de referencia para los debates sobre ética de la IA, mientras que, para otros, carecen de una orientación específica sobre cómo aplicarlos a la práctica (Peller, 2023). En todo caso, la carta de The Future of Life Institute, conforme a estos principios, proponía que la IA avanzada debía ser planificada y gestionarse con el cuidado y los recursos adecuados. Instaba a supeditar su desarrollo a la certeza de que sus efectos serán positivos y sus riesgos controlables.



Recapitulando, el diseño y desarrollo de la IA por parte de las *big tech* refleja lógicas sociales, políticas y económicas concretas, que configuran el discurso dominante, según el cual este proceso compromete un "proyecto civilizatorio orientado continuamente, y en todas partes, a organizar el mejor avance del mundo" (Sadin, 2024, p. 25). De lo expuesto, se advierte que esta forma de definir el problema ya prefigura en su enunciado a los potenciales ganadores y perdedores.

Para Innerarity (2025), la problematización sobre la IA suele reducirse a posturas extremas y polarizadas. Por un lado, las optimistas que ven en la tecnología un destino inevitable con el poder de sustituir a aquellas instituciones debilitadas e ineficientes, y por otro, las centradas en los peligros de la IA que la culpan de erosionar la gobernanza democrática. Ambos posicionamientos parten del supuesto de que la tecnología puede reemplazar a la política. En este punto, se manifiesta una tensión histórica inherente a la relación entre el conocimiento experto y la esfera política: entre la objetividad técnica fundamentada en la precisión de sus procedimientos y la preocupación por la despolitización tecnocrática. Este debate ha abonado la extendida creencia social en la neutralidad de la tecnología. No obstante, tras el análisis previo, resulta posible afirmar que la IA trasciende los asuntos técnicos o la consideración exclusiva de la inteligencia, y que carece de toda neutralidad en términos de política y poder (Coeckelbergh, 2022).

A lo anterior, se adiciona que estas posturas se han construido en gran medida sobre la base de discusiones desplegadas en el Norte Global, atendiendo a problemas

que le son propios, con capacidades y regulaciones preexistentes e institucionalidades generalmente más robustas que las de los países de América Latina. Al colocar el foco en la Unión Europea, se advierte que el problema se define por las probabilidades de peligro y la "agenda de los riesgos". Es decir, cómo se valoran los riesgos, cuáles son aceptables y qué conductas se recomiendan en consecuencia (Innerarity, 2013, p. 69). En este sentido, es preciso no perder de vista que toda medida preventiva conlleva riesgos, tanto por lo que se hace como por lo que se omite. Luego de un proceso cargado de negociaciones, marchas y contramarchas, se presentó el primer marco jurídico a nivel global sobre el tema, la Ley de Inteligencia Artificial, aprobada por el Parlamento Europeo el 13 de marzo de 2024 y de implementación gradual.

El *AI Act* procura garantizar que los sistemas de IA utilizados en esta región sean seguros, transparentes, trazables y responsables, no discriminatorios y respetuosos con el medioambiente. Lo distintivo es que establece obligaciones para proveedores y usuarios en función del nivel de riesgo de la IA involucrada. De acuerdo con la clasificación que reciben, se aplican diferentes abordajes regulatorios: desde inaceptable a aceptable con diversos matices. La legislación incluye también un apartado específico sobre la IA generativa, que contiene varios requisitos de transparencia como, por ejemplo, que el sistema debe revelar cuando un contenido ha sido generado por IA, diseñar el modelo para evitar que genere contenidos ilegales y publicar resúmenes de los datos protegidos por derechos de autor utilizados para el entrenamiento. De este modo, la norma abre un frente interesante para la discusión sobre los datos fundacionales para el entrenamiento de grandes modelos de lenguaje basados en IA (Peller, 2023).



La experiencia de la Unión Europea ofrece un marco de referencia valioso para los debates en desarrollo en América Latina. No obstante, resulta fundamental evitar la mera extrapolación o "importación" de problemas, agendas y marcos regulatorios. En consonancia con Innerarity, esto demanda "el retorno de la política como revitalización del Estado, recuperación de la lógica política y gestión democrática de los riesgos" (2013, p. 73), aun cuando los actores involucrados carezcan de los instrumentos necesarios para ejercer plenamente dicha responsabilidad. La forma en la que se definen los problemas de la IA y las agendas en cada región presenta particularidades, más allá de sus puntos en común. De allí que no se considere apropiado hablar de problemas de la IA en general o de manera homogénea, sino desde su construcción social en determinados momentos históricos y geográficos, a partir de tensiones y consensos que irán variando a lo largo del proceso.

## 2. El atravesamiento de las vidas y la cotidianidad

La IA está presente en casi todos los aspectos de la vida cotidiana; sin embargo, el acceso y sus formas de adopción no pueden considerarse un problema de carácter privado de cada usuario. Sadin plantea que el avance de la digitalización está instituyendo "un acompañamiento algorítmico de la vida" (2024, p. 21), donde los flujos de la vitales se confunden con los digitales. Según este autor, la historia de la técnica

es indisociable de la historia del cuerpo, lo que lleva a reflexionar sobre cómo las tecnologías moldean y redefinen las capacidades, límites y experiencias corporales, lo cual se ha visto potenciado a partir de la interconectividad global –personas, sistemas, dispositivos– en tiempo real.

Con el inicio de la "era del acceso" y la masificación de internet en el último tercio del 1990, fue emergiendo y consolidándose una arquitectura tecnológica respaldada por una nueva economía que llegaba para facilitar la vida. A mediados de la década siguiente, aparecieron nuevas funcionalidades que permitieron a las personas no solo consultar documentos sino también intervenirlos, pasando de ser meros espectadores a potenciales productores de datos e información. Esto se expandió con las redes sociales y el advenimiento de la web 2.0. El primer punto de inflexión se produjo en 2007 con el lanzamiento del IPhone, que inauguró una lógica ético-económica nueva: la abolición, a gran velocidad, de la brecha que separa a uno de otro, a cada ser o cada cosa, en favor de la recomendación personalizada y automatizada de productos y servicios con miras a garantizar un supuesto mejor desarrollo de la vida cotidiana (Sadin, 2024). En noviembre de 2022, se identifica un segundo punto de inflexión con el lanzamiento global de ChatGPT y su capacidad de generar contenidos nuevos y originales a partir de datos existentes.[2]

Retomando el concepto de Sadin (2024, pp. 20-23), todos estos avances tecnológicos comparten una característica fundamental: la capacidad de orientar discretamente las elecciones humanas con fines interesados. Así, se consolida lo que el autor denomina la era del "integralitarismo digital", donde cada aspecto de la vida queda sometido a una mercantilización continua. En este contexto, incluso las interacciones más íntimas se transforman en datos comercializables. Uno de los ejemplos más reveladores de esta adopción acrítica de la tecnología es la construcción de la huella digital desde el nacimiento. Popularizado como *sharenting*,[3] refiere a la práctica de compartir información de niños sin considerar las implicaciones en la identidad digital y privacidad de esta exposición temprana que incluye, entre otras cuestiones, el uso del contenido que estos usuarios publican para entrenar modelos de IA generativa.[4]



La expansión del discurso dominante basado en la innovación tecnológica y sus atributos para mejorar la vida, ha contribuido a fomentar la idea de que la tecnología puede autorregularse. Pedace *et al.* plantean que considerar que el problema creado por una tecnología será compensado por otra o resuelto por la misma tecnología podría

[2] De acuerdo con el *AI Act* de la Unión Europea, la IA generativa es un modelo de base utilizado en un sistema de IA destinado específicamente a generar contenidos con distintos niveles de autonomía, como texto complejo, imágenes, audio o vídeo (Lamm *et al*., 2023).

[3] Para ampliar se recomienda la siguiente lectura: https://www.unicef.org/parenting/es/salud-mental/sharenting-compartir-informacion-sobre-hijos-en-linea.

[4] En junio de 2024, la empresa Meta anunció una nueva política de privacidad que habilitaba el uso del contenido público de usuarios en sus plataformas para entrenar modelos de IA generativa. Aunque la compañía afirmó que únicamente procesaría información pública, surgieron diversas críticas por la falta de transparencia (Cantero Gamito & Bosoer, 2025).

conducir a una dependencia excesiva de la IA como solución a las preocupaciones éticas: "Si los problemas se consideran puramente tecnológicos, sólo deberían requerir soluciones tecnológicas. En lugar de eso, tenemos decisiones humanas disfrazadas de tecnología" (2023, p. 4).

Si bien los debates sobre innovación responsable y diseño tecnológico sensible a los valores no son nuevos, en la práctica estas decisiones están concentradas. Coeckelbergh (2022) argumenta que la libertad no debería limitarse la capacidad de elección del usuario, sino que debe incluir la posibilidad de participar activamente en los procesos de innovación y regulación de las tecnologías que los afectan.[5] Esto es crucial ya que la tecnología no solo tiene efectos no intencionados, sino que moldea las sociedades, las vidas y las subjetividades. Según este autor, en ausencia de esta forma de libertad como participación, el futuro de la IA quedará en manos de tecnócratas y los CEO de las grandes corporaciones.

### 3. Convergencias tecnológicas y asimetrías de poder en los datos

Para los sistemas de IA, los datos y los algoritmos son complementarios e igualmente indispensables. En la medida en que estos modelos se vuelven mejores con el uso, pueden ser considerados como "medios de producción", con la particularidad de que, en lugar de depreciarse, se aprecian con el tiempo (Rikap, 2023). Siguiendo a esta autora, las actuales técnicas –como el aprendizaje profundo o *deep learning*–, se perfeccionan cuantos más datos procesan, lo que externaliza parte de su optimización en los usuarios cada vez que estos realizan una búsqueda, una pregunta, o publican contenidos en redes sociales.



En sociedades donde predomina la cultura del compartir, las personas generan una gran cantidad de datos georreferenciados que registran sus diversos movimientos. Esta práctica adquiere una dimensión problemática cuando se combina con la falta o poca transparencia en el funcionamiento de los sistemas digitales. Un ejemplo cotidiano de esta opacidad se manifiesta en la aceptación acrítica de los términos y condiciones de aplicaciones y redes sociales, que a menudo son aprobados sin una lectura detenida por parte de los usuarios, mientras que las empresas analizan estos comportamientos y los utilizan para optimizar sus estrategias comerciales. Si bien se podría argumentar que la información sobre el uso de datos está disponible, es poco accesible en términos de comprensión de sus implicancias para el usuario promedio.

[5] El autor comienza el recorrido desde la definición de libertad de los antiguos, para quienes la libertad no estaba planteada en términos de elección sino de acción política (Arendt, 1958). Una expresión moderna, es la libertad como participación en la toma de decisiones políticas de Rousseau, quien argumentaba ya antes que Kant que la libertad significa autorregularse. Rousseau también otorgó a la autorregulación un significado político: si los ciudadanos crean sus propias leyes, son realmente libres y no se encuentran a merced de la tiranía de otros (Coeckelbergh, 2022, p. 40).

En este punto, Mantegna (2022) argumenta que la opacidad tecnológica puede manifestarse de dos maneras: en términos técnicos, cuando la complejidad algorítmica dificulta la comprensión de su funcionamiento por parte de quienes carecen de conocimientos especializados; y en términos legales, cuando el acceso a la información está restringido por dispositivos jurídicos que convierten los sistemas en "cajas negras", impidiendo su auditoría y control ciudadano. Esto se vincula, al mismo tiempo, con la dificultad –e incluso imposibilidad– de comprender cómo un sistema llega a determinados resultados, aun para los propios desarrolladores o expertos en la materia, lo que se denomina "inescrutabilidad algorítmica".

La pandemia de COVID-19 aceleró estas prácticas, y la urgencia tendió a primar sobre las cuestiones éticas relacionadas con la gobernanza de datos. Esta lógica ha contribuido al incremento en las asimetrías de poder, "generando dinámicas desfavorables a las personas usuarias y dueñas de los datos" (Lamm *et al.*, 2023, p. 63). Solo para poner un ejemplo, cada vez que se utiliza un sistema de recomendación basado en IA, sus anuncios altamente personalizados influyen en el comportamiento del consumidor en la dirección sugerida por el algoritmo. Este proceso -que, *a priori*, no implica una coerción directa- opera a través de lo que Thaler y Sunstein (2009, en Coeckelbergh, 2022) denominan *nudge* o "suave empujón", un mecanismo que permea en los deseos y elecciones individuales bajo la premisa de orientar al usuario hacia su mejor interés. Estos sistemas funcionan en el umbral entre la persuasión y la manipulación, y se potencian mediante la convergencia de tecnologías digitales.



Ante el avance en materia de tecnologías inmersivas impulsado por la IA generativa, Sadin introduce lo que considera la fase final del acompañamiento algorítmico de la vida, "la encapsulación de los cuerpos y las mentes dentro de un entorno técnico-económico que abarca prácticamente la totalidad de la experiencia cotidiana y los procesos cognitivos" (2024, pp. 33-34). Esta etapa representa un nivel inédito de imbricación entre los seres humanos y la tecnología, facilitado por el desarrollo de las neurotecnologías. Su rasgo distintivo es la capacidad de establecer una interfaz directa con el cerebro, influenciando tanto la percepción como la toma de decisiones. En palabras de Mantegna (2024), la futura evolución de internet se encuentra en el metaverso, una intersección de tecnologías de realidad extendida, que utilizan distintas aplicaciones de IA y cuya forma de acceso involucrará dispositivos de *hardware* con capacidades neurotecnológicas. De este modo, en un ecosistema donde los datos personales se tratan como bienes intercambiables, el uso comercial de datos neuronales o cerebrales ha abierto diversos debates éticos y, en consecuencia, jurídicos.

Ahora bien, previo a abordar estas discusiones, es preciso profundizar en algunos términos. Por neurotecnología se entiende a todo dispositivo y procedimiento utilizado para acceder, monitorear, investigar, evaluar, manipular y emular la estructura y función de los sistemas neuronales. Se incluyen aquí herramientas técnicas y computacionales que miden y analizan señales químicas y eléctricas en el sistema nervioso, así como aquellas que interactúan con el sistema nervioso para cambiar su actividad y restaurar la información sensorial, como por ejemplo la estimulación cerebral profunda para

detener la vibración y otras condiciones patológicas (UNESCO & IBC, 2021, en Lamm *et al.*, 2023). En este punto, es importante distinguir entre aquellos dispositivos invasivos que requieren implantación quirúrgica (de uso exclusivamente médico), y aquellos no invasivos de uso comercial como cascos o sensores externos. Estos últimos, también llamados neurotecnologías vestibles o *wearables*, están en pleno desarrollo liderado por empresas como como Neuralink, Meta, Snap y Apple.

El primer hito en el campo de las neurotecnologías se registró en 1878, cuando Richard Canton descubrió la transmisión de señales eléctricas a través del cerebro de un animal. Casi cincuenta años después, se realizó la primera electroencefalografía en humanos. En la década de 1990, el desarrollo de la resonancia magnética funcional permitió medir la actividad eléctrica del cerebro de forma indirecta utilizando el flujo sanguíneo. Esto posibilitó cartografiar el funcionamiento del cerebro y fue un avance sin precedentes en términos de diagnósticos preventivos o terapéuticos. En los últimos años, se ha demostrado que estas tecnologías también son eficaces para conocer las intenciones, opiniones y actitudes de las personas. Los escáneres cerebrales no solo permiten leer intenciones y recuerdos concretos relacionados con los experimentos, sino que incluso parecen capaces de descodificar preferencias más generales como hábitos de consumo u opiniones políticas (Ienca & Andorno, 2017).

El ritmo acelerado de la investigación en neurociencias y sus tecnologías asociadas ha suscitado preocupaciones sobre sus potenciales impactos y la vulneración de derechos fundamentales. El entonces presidente de los Estados Unidos, Barack Obama, fue uno de los primeros actores en poner en la agenda pública la necesidad de abordar estos desafíos, destacando aspectos como la integridad mental, la privacidad y la autonomía de los individuos (Presidential Commission for the Study of Bioethical Issues, 2014). En 2023 se firmó la declaración europea para proteger los derechos digitales frente al desarrollo de la neurotecnología, en particular aquella de uso comercial, lo que se conoció como Declaración de León. En este documento se alerta sobre la posibilidad de que los datos cerebrales sean acumulados y comercializados por un número reducido de empresas -tendencia registrada en la actualidad-, dejando latentes riesgos de ciberataques, desinformación y manipulación del pensamiento (Declaración de León, 2023, p. 2). Aspectos similares se discutieron en la Conferencia Internacional sobre la Ética de la Neurotecnología (UNESCO, 2023), donde se problematizaron las diferencias en términos de regulación ética y jurídica entre las neurotecnologías de uso médico y las de uso comercial.



Esta cuestión puede ser abordada desde diversas aristas. Coeckelbergh (2022) señala el peligro de considerar a las personas como objetos susceptibles de ser manipulados, ya sea con el argumento de su propio bien o por el bien de la sociedad. Autores como Ienca y Andorno (2017) advierten que el avance sin regulación adecuada podría permitir formas de manipulación sin precedentes en la esfera privada, pudiendo causar daños físicos o psicológicos, o permitiendo una influencia indebida en su comportamiento que afecte directamente la dignidad humana. Por su parte, Mantegna (2024) puntualiza en los riesgos asociados a la reidentificación de datos cerebrales

mediante la reversión del proceso de anonimización, y en la seguridad informática ante posibles ciberataques. Es decir, el riesgo de piratería de los datos cerebrales, su reutilización y comercialización no autorizada. En este punto, la autora enfatiza las diversas formas de vigilancia digital que pueden desplegarse a partir de inferencias extraídas de dichos datos, así como la recopilación y el tratamiento de datos cerebrales para fines no consentidos por la persona.

Tal como se observa, las neurotecnologías están estrechamente vinculadas con el desarrollo de la IA y comparten riesgos, por lo que no pueden ser tratadas como conversaciones separadas (CLAD, 2023). En gran medida, su problematización ha girado en torno a la necesidad de que la información mental inferida de los datos cerebrales tenga una protección especial, que la distinga de otros tipos de datos personales. Un antecedente de esta discusión es la regulación del acceso a los datos genéticos humanos. La Declaración Universal sobre el Genoma Humano y los Derechos Humanos (1997) procura evitar que la información genética se recopile y utilice de forma incompatible con el respeto de los derechos fundamentales, y proteger el genoma humano de manipulaciones indebidas que puedan perjudicar a las generaciones futuras. Posteriormente, la Declaración Internacional sobre los Datos Genéticos Humanos (2003) estableció normas más específicas para la recogida de muestras biológicas humanas y datos genéticos. En este marco, se integraron nuevos derechos, como el de decidir sobre conocer la propia información genética, y se adaptaron derechos ya normados a los retos que planteaba el desarrollo de la genética, como el derecho a la intimidad y el derecho contra la discriminación (Ienca & Andorno, 2017).



En la actualidad, la comunidad internacional se encuentra en una situación similar: la necesidad de revisar no solo nociones jurídicas, sino también éticas, en las investigaciones en neurociencias y sus tecnologías asociadas. Por un lado, son cada vez más las iniciativas que reconocen la importancia de preservar la libertad cognitiva, o al menos consideran preciso un marco de gobernanza específico para su protección, mientras por otro lado se plantea que los marcos existentes y leyes de protección de datos personales pueden ser aplicados también a la protección de datos cerebrales (Cantero Gamito, 2024). En este panorama, el enfoque de los neuroderechos se ha instalado como una respuesta a los desafíos que plantea el desarrollo de las neurotecnologías, especialmente las de uso comercial.

El derecho neuronal o neuroderecho hace referencia a “los principios éticos, jurídicos, sociales o naturales de libertad o derecho relacionados con el dominio cerebral y mental de una persona; es decir, las reglas normativas fundamentales para la protección y preservación del cerebro y la mente humanos” (Ienca, 2021, p. 1). Entre estos principios, se menciona: la identidad personal entendida como la protección de la integridad física y mental que garantice que nadie pueda leer, alterar o difundir dicha información sin su consentimiento; también el principio de autonomía y libre toma de decisiones; y la privacidad mental en cuanto recopilación y procesamiento de datos de neurodispositivos que pueden potencialmente identificar a las personas, relevar su actividad cerebral y dar lugar a prácticas discriminatorias. En este punto,

se resalta como crítico el acceso equitativo a estas tecnologías y la protección de los usuarios frente a posibles sesgos algorítmicos (Andorno, 2023; Cantero Gamito & Bosoer, 2025).

A pesar de que el *mainstream* está concentrado en el Norte Global, diversos países de América Latina han priorizado este debate en sus agendas públicas. El caso más destacado es el de Chile que, en 2021, fue pionero en reconocer a nivel constitucional los derechos humanos en relación con la neurotecnología. En agosto de 2023, este país volvió a ocupar un lugar central en la discusión global, cuando su Corte Suprema admitió una acción de protección interpuesta por un usuario contra la empresa Emotiv Inc., en relación con el dispositivo Insight comercializado por esta firma. Se trata de un dispositivo diseñado para la captación de datos neuronales que opera como una vincha equipada con sensores capaces de registrar la actividad eléctrica del cerebro y extraer información sobre gestos, movimientos, preferencias, tiempos de reacción y actividad cognitiva del usuario (Siverino Bavio, 2025, pp. 54-55).

El poder judicial determinó que la empresa había vulnerado garantías constitucionales fundamentales como el derecho a la integridad física y psíquica y el derecho a la privacidad. La sentencia enfatizó dos aspectos: i) la falta de consentimiento y protección de los datos neuronales del demandante, ante su reclamó por la eliminación gratuita de los datos generados a través del uso del dispositivo; y ii) la comercialización de una tecnología sin evaluación previa de sus impactos en los derechos humanos. Consecuentemente, la Corte Suprema ordenó a la empresa la eliminación inmediata de toda la información almacenada en su nube y portales digitales relacionada con el demandante. El caso Girardi *vs.* Emotiv, constituye el primer antecedente a nivel mundial en materia de protección de datos neuronales y neuroderechos. Según Cantero Gamito (2024), esta sentencia sienta un precedente en la regulación de neurotecnologías y su compatibilidad con la protección de la vida privada. Asimismo, el fallo subraya la urgencia de establecer controles rigurosos sobre los dispositivos de uso comercial, antes de que sean introducidos en el mercado (Mantegna, 2024).



Este ejemplo demuestra la importancia de tratar a los datos neuronales como datos sensibles. La Declaración de Principios Interamericanos sobre Neurociencias, Neurotecnologías y Derechos Humanos (2023) incluye estándares específicos relacionados con la protección de datos personales y, en general, de los derechos de las personas desde la etapa de diseño de las neurotecnologías hasta su implementación, evaluación, comercialización y uso (Monti, 2025). En este sentido, se plantea que el principio de consentimiento informado podría resultar insuficiente en la adopción de este tipo de tecnologías, ya que supone que la persona conoce y comprende plenamente el tratamiento que se hará de sus datos. En palabras de Cantero Gamito y Bosoer (2025), muchas empresas prefieren utilizan un sistema de *opt-out*, donde los usuarios deben rechazar activamente el uso de sus datos, por ejemplo, para evitar recibir publicidad personalizada; en lugar de un *opt-in* o aceptación, que permita a los usuarios decidir activamente si quieren que sus datos sean utilizados de esta forma.

Es preciso considerar que los datos recopilados con un propósito hoy pueden ser reutilizados en el futuro para fines completamente diferentes, a menudo sin el conocimiento o la aprobación explícita de los usuarios. Por ello, se vuelve vital revisar los marcos regulatorios a fin de que permitan a las personas actualizar y gestionar continuamente sus preferencias sobre cómo se utilizan sus datos. Esto se conjuga en el emergente concepto de "consentimiento informado dinámico" (Cantero Gamito & Bosoer, 2025, p. 52).

Todas las problemáticas arriba mencionadas ponen de manifiesto la complejidad inherente al desarrollo científico-tecnológico contemporáneo, así como las asimetrías de poder que atraviesan el ciclo de vida de las tecnologías. Al mismo tiempo, que expresan, en diversos debates, las preocupaciones crecientes en torno a identidad, privacidad e integridad humana, en un escenario marcado por la concentración de la producción del conocimiento.

## 4. La democratización del conocimiento como una cuestión ética

En las "realidades pixeladas", el aumento de la influencia de la industria de datos, las redes y las IA está llevando al debilitamiento de las categorías políticas y morales que hasta hace poco se consideraban intocables (Sadin, 2024, p. 30). En otras palabras, los sistemas de IA, principalmente a través de su integración en las redes sociales, están transformando las bases epistémicas de la sociedad (Coeckelbergh, 2022). La porosidad de los espacios de deliberación pública en Internet es cada vez más preocupante y es un reflejo, tal como se evidenció en el punto anterior, de la pérdida progresiva de la autonomía de las personas en sus cuerpos y en sus mentes.



Existe un consenso creciente en torno a la idea de que los sistemas de IA reflejan y reproducen los valores, normas y estructuras sociales dominantes. Las *big tech* no solo monopolizan el desarrollo de la IA más avanzada; son también las dueñas de las mayores bases de datos que están en continua expansión a medida que se utilizan sus plataformas e infraestructura digital (Rikap, 2023). Los sistemas de IA, diseñados bajo la lógica del capitalismo de plataformas, tienden a la homogeneización de conductas y discursos. Dicho de otro modo, se corre el riesgo de que estas tecnologías refuercen una "cultura uniforme donde los valores más visibles sean la eficiencia, la optimización, la rapidez, la productividad y el éxito. Se trata de una cultura que descansa sobre un conjunto específico de temas, y excluye o limita la visibilidad de otros" (Cantero Gamito & Bosoer, 2025, p. 26).

Lo anterior retrotrae el tiempo a una de las primeras problematizaciones sobre los sistemas de IA, aquella que plantea que el diseño y desarrollo de estas tecnologías "ocultan" e intensifican asimetrías y discriminaciones históricas. Para entender esta cuestión, es preciso hacer alusión al concepto de sesgo algorítmico, definido como las anomalías en los resultados generados por la IA, producto tanto de los datos utilizados en la fase de entrenamiento como de los sesgos presentes en el diseño y desarrollo

del sistema. En gran medida, estas distorsiones son fruto lo que se denomina "riesgo del privilegio"; es decir, la dificultad de las personas en posiciones de poder para reconocer las dinámicas de opresión existentes (D'Ignazio & Klein, 2023, p. 10). Para estas autoras, el problema radica en la composición homogénea de quienes diseñan y desarrollan los sistemas. La industria de la IA está dominada predominantemente por hombres blancos de élite, lo que implica un sesgo de privilegio colectivo que limita la capacidad de identificar y mitigar discriminaciones antes de que los algoritmos sean desplegados en el mundo (D'Ignazio & Klein, 2023).

En América Latina, este problema se agrava debido a que los sistemas de IA de uso masivo se entrenan, salvo contadas excepciones, con bases de datos que no representan la diversidad poblacional, ni las costumbres, ni las características locales. Este fenómeno, conocido como "colonialismo de datos", se basa en la adopción de tecnologías diseñadas en contextos ajenos, con flujos de datos poco regulados y asociados a una infraestructura tecnológica dependiente (Filgueiras, 2023, p. 66).

Siguiendo a Breilh (2022), el control monopólico de la información y la producción del conocimiento científico restringe el acceso equitativo a una ciencia veraz, confiable y democrática. En este contexto, la crisis ética del conocimiento se sustenta en la ignorancia estratégica y la desinformación planificada, aspectos que configuran una "infodemia profunda". Los modelos algorítmicos en su búsqueda de la eficiencia generan respuestas basadas en patrones previamente aprendidos, incluso cuando estos contienen información falsa o discriminatoria.. A nivel colectivo, esto refuerza los sesgos cognitivos de los usuarios, que ven confirmado lo que ya pensaban, contribuyendo así a la radicalización de sus opiniones. La diseminación de las "burbujas de filtro", mediante las cuales los algoritmos amplifican una visión homogénea de la realidad, actúa como una caja de resonancia que aísla del disenso (Mantegna, 2022).



Sobre esta base, cualquier esfuerzo por mitigar los sesgos algorítmicos debe comenzar por una revisión crítica de los criterios con que se construyen las bases de datos, a fin de contrarrestar los procesos sistemáticos de "invisibilización" de colectivos. Como señalan D'Ignazio y Klein (2023), un primer paso es el reconocimiento de la brecha de género en los datos –*gender data gap*–, un fenómeno que evidencia cómo la investigación científica y tecnológica ha estado históricamente sesgada por el predominio de datos extraídos de cuerpos de varones. Este problema se profundiza cuando la falta de información desagregada y diversa limita la toma de decisiones en materia de políticas públicas. En otras palabras, la recolección de datos que no contempla la interseccionalidad –como la raza, el género o el nivel socioeconómico– contribuye a la exclusión sistemática de grupos vulnerables. Esta carencia u omisión en la construcción de los datos se vuelve un riesgo al facilitar su uso con fines discriminatorios, perpetuando y profundizando las desigualdades existentes.

Lo anterior se vincula, en gran parte, con los modos en que se generan y apropian los conocimientos científicos y tecnológicos en la actualidad. Rikap (2023, p. 74) llama la atención sobre la progresiva concentración del *expertise* en las *big*

*tech* estadounidenses. Otra práctica frecuente es la contratación de investigadores e investigadoras referentes en universidades de diversas partes del mundo, "quienes ofician de puente entre la coproducción de IA con el mundo académico y la apropiación de resultados por parte de las empresas privadas", principalmente a través de las patentes. Por consiguiente, la concentración de la información y el conocimiento, el colonialismo de datos, la privatización de la ciencia y la tecnología, y la desinformación planificada se presentan como diversas formas de vulneración al derecho a la ciencia.

Este derecho, reconocido como un derecho cultural fundamental, se sostiene para su pleno ejercicio en el acceso universal y sin discriminación a los beneficios de la ciencia –básica y aplicada–, así como a oportunidades equitativas para contribuir a la actividad científica y la garantía de libertad en la investigación, la participación en la toma de decisiones sobre el desarrollo científico y tecnológico, y un entorno favorable para la conservación, el desarrollo y la difusión del conocimiento científico y tecnológico (CDESC, 2020; Relatora Especial, 2012, citado en Bohoslavsky, 2022).

En este panorama, que a primera vista puede parecer desalentador, Rikap (2025) vislumbra una ventana de oportunidad para desarrollar sistemas de IA "pública" que operen fuera de la lógica geopolítica imperante. En otras palabras, alude a sistemas públicos, abiertos y auditables, diseñados para fortalecer la participación democrática. Esta perspectiva disputa las definiciones dominantes del problema, con foco en el uso y el usuario, y propone su reformulación hacia construcciones colectivas y colaborativas. Al mismo tiempo, reposiciona al Estado y a lo público en un rol central, en cuanto a la generación de capacidades institucionales para la definición soberana de la agenda de desarrollo digital.



Esta labor incluye a los tres poderes del Estado, tanto en la priorización de cuestiones y su despliegue en políticas públicas –como un marco de gobernanza de la IA y su regulación–, como en el cumplimiento de las legislaciones nacionales e internacionales, la supervisión de la esfera digital pública y la rendición de cuentas de las empresas. En este último punto, uno de los casos más difundidos en la región es el de Brasil y la prohibición de la red social X, tras reiterados incumplimientos de las órdenes judiciales que le exigían a la empresa bloquear aquellos perfiles dedicados a la divulgación de discursos discriminatorios, de odio y antidemocráticos.[6]

En su concepción amplia, el derecho a la ciencia no solo protege la labor de las y los científicos, sino que también incorpora una dimensión social que incluye a todas las personas que, sin ser especialistas, se ven involucradas o afectadas por el desarrollo científico y tecnológico (Saba, 2021, citado en Bohoslavsky, 2022). En particular, se han ampliado los debates en torno a la introducción de la IA en la educación formal y las oportunidades de alfabetización científica y tecnológica para la población en general.

[6] Para ampliar esta información, se recomienda la siguiente lectura: https://elpais.com/tecnologia/2024-09-17/necesitamos-reclamar-nuestra-soberania-digital-una-carta-de-intelectuales-denuncia-presiones-de-las-tecnologicas-a-brasil.html.

En este sentido, la urgencia radica en la necesidad de estimular habilidades digitales que permitan a la ciudadanía "abrir las cajas negras" de los sistemas de IA. Según Collett *et al.* (2022, p. 40), se trata de fomentar la capacidad de gestionar, comprender, integrar, comunicar, evaluar y crear información de forma segura y apropiada a través de dispositivos digitales y tecnologías de red para la participación en la vida económica y social. Queda claro una vez más que, dada la creciente complejidad de estas tecnologías, la estrategia para su abordaje debe ser integral e involucrar responsablemente a todos los actores que participan en los diversos momentos del ciclo de vida de las IA.

La democratización del conocimiento científico y tecnológico requiere mecanismos de transparencia, supervisión y participación ciudadana, evitando caer en prácticas de *ethics washing*. Es decir, la adopción de principios éticos por parte de empresas tecnológicas o gobiernos como un recurso superficial o simbólico para mejorar su legitimidad, pero sin un compromiso efectivo con políticas y cambios concretos en sus prácticas (Cantero Gamito & Bosoer, 2025). Siguiendo a estas autoras, el colonialismo de datos y la dependencia tecnológica se acentúa cuando se adoptan marcos éticos que tienen un carácter universalista, y se fundamentan en concepciones utilitaristas e individualistas que no necesariamente reflejan la diversidad de valores y realidades culturales. Los mismos suelen ser diseñados desde perspectivas eurocéntricas, asumen estándares homogéneos aplicables a nivel global e ignoran los legados históricos y culturales de cada región del planeta. Por ejemplo, la idea de una IA "centrada en el ser humano", tan frecuente en los debates éticos globales, "puede no resonar con comunidades indígenas que valoran el colectivo sobre el individuo o que entienden la tecnología desde una relación más integral con la naturaleza" (Cantero Gamito & Bosoer, 2025, p. 27).



Recapitulando, la construcción de marcos éticos genuinamente inclusivos y representativos requiere de pluralidad epistémica (Bohoslavsky, 2022), que reconozca la diversidad de valores en cada contexto histórico, social y geográfico, como un modo de contrarrestar los discursos dominantes y homogeneizadores.

## Reflexiones finales

A lo largo de este artículo, se ha planteado la importancia de entender cómo y quiénes definen los "problemas de la IA", ya que ello indefectiblemente envuelve valores, posicionamientos e intereses particulares. La priorización de "una IA que reemplaza en lugar de potenciar" –a las personas–, así como el foco puesto en la regulación del uso –y no de su producción–, es resultado de decisiones económicas y políticas de las *big tech* (Rikap, 2023). Los actores que concentran el diseño y desarrollo de las IA tienen el poder de diseminar sus discursos, profundizando en este proceso las desigualdades preexistentes. Se trata también de poner en valor la ciencia y la tecnología como bienes públicos ante su excesiva privatización: "Es urgente debatir democráticamente qué IA queremos como sociedad y para qué. Este debate es indisociable de la pregunta sobre qué datos han de ser recopilados -y cuáles no- y cómo se decidirá quién accede y cómo a las bases resultantes" (Rikap, 2023, p. 82).

En línea con lo anterior, Gómez Mont *et al.* (2024) plantean la necesidad de una IA para el bien social que apunte en la dirección del empoderamiento de las personas, mientras que Innerarity va un paso más allá, al proponer una "política de la humanidad" tendiente a reducir las asimetrías entre los que deciden y los que padecen: "No hace falta que nosotros seamos todos (algo que no es ni posible ni bueno), pero deberíamos mantener siempre activa la pregunta de si somos todos los que estamos y estamos todos los que somos" (2013, p. 208).

Así como la definición de los problemas y las decisiones sobre cuál de estas cuestiones se prioriza en cada agenda no son neutrales, tampoco lo son los sistemas de IA. Su diseño, implementación y potenciales impactos varían según los contextos históricos y políticos en los que se desarrollan y utilizan. Los llamados riesgos de las IA no son homogéneos, ya que reproducen y amplifican las dinámicas de poder que circulan en las sociedades. La construcción de capacidades políticas es crucial para encauzar el desarrollo tecnológico de manera ética y democrática. Esta problemática no se limita a las decisiones de las grandes empresas tecnológicas, sino que involucra directamente a los gobiernos cuando hacen uso acrítico –o al límite de la vulneración de derechos– de las tecnologías digitales, desplegando estrategias de control y vigilancia de la ciudadanía.

En las páginas anteriores se ha argumentado que la IA debe ser comprendida como una construcción social profundamente atravesada por la dimensión política. En este marco, la reivindicación del derecho a la ciencia implica, por un lado, la reducción de las brechas de acceso y la promoción de un uso equitativo y no discriminatorio de la tecnología, y por otro exige la participación activa en la toma de decisiones sobre el desarrollo científico y tecnológico, un proceso clave en la creación de entornos que favorezcan la democratización del conocimiento. A fin de cuentas, los problemas de la IA no son meramente técnicos sino fundamentalmente políticos, y requieren de debates que territorialicen las necesidades y los conflictos, atendiendo las particularidades de cada sociedad.



El discurso de la inevitabilidad y urgencia del desarrollo tecnológico para el progreso de la humanidad esconde los esfuerzos del capitalismo de plataformas por homogeneizar las culturas. Por el contrario, su trayectoria puede ser moldeada por decisiones colectivas que prioricen el bienestar social por sobre la lógica de acumulación. Las respuestas deben orientarse a la construcción de ciencia y tecnología diversa, democrática y accesible, junto con la formulación de marcos éticos genuinamente inclusivos, capaces de representar la pluralidad epistémica. Marcos éticos que sirvan como herramientas para la reconfiguración de las "realidades pixeladas".

## Reconocimiento de uso de IA

Este artículo es resultado del trabajo de indagación y reflexión crítica de la autora. En su elaboración, se han utilizado sistemas de IA (ChatGPT y Gemini), para

corroborar normas de estilo, realizar correcciones gramaticales, y en particular como herramientas de apoyo en la elaboración y traducción de los resúmenes y de las referencias bibliográficas en lengua inglesa.

**Bibliografía**

Aguilar Villanueva, L. (1993). Problemas públicos y agenda de gobierno. México: Porrúa Grupo Editorial.

Andorno, R. (2023). Neurotecnologías y derechos humanos en América Latina y el Caribe: Desafíos y propuestas de política pública. París: UNESCO. Recuperado de: https://unesdoc.unesco.org/ark:/48223/pf0000387079.

Bertoni, E. (2024). Tecnologías y derechos humanos. En M. Moisés Sánchez, C. Colombara & N. Monti (Eds.), En defensa de los neuroderechos (28-33). Santiago de Chile: KAMANAU. Recuperado de: https://defensaneuroderechos.org.

Bohoslavsky, J. (2022). Pluralidad epistémica y derechos humanos en pandemia. En J. Bohoslavsky (Comp.), Ciencias y pandemia: una epistemología para los derechos humanos (19-44). La Plata: EDULP. Recuperado de: https://libros.unlp.edu.ar/index.php/unlp/catalog/book/1920.



Breilh, J. (2022). La pandemia y el derecho a una ciencia veraz, humilde y emancipadora. En J. Bohoslavsky (Comp.), Ciencias y pandemia: una epistemología para los derechos humanos (12-18). La Plata: EDULP. Recuperado de: https://libros.unlp.edu.ar/index.php/unlp/catalog/book/1920.

Cantero Gamito, M. (2024). Neurotecnología y privacidad: Chile lidera el cambio legal. En M. Moisés Sánchez, C. Colombara & N. Monti (Eds.), En defensa de los neuroderechos (42-47). Santiago de Chile: KAMANAU. Recuperado de: https://defensaneuroderechos.org.

Cantero Gamito, M. & Bosoer, L. (2025). Desafíos estructurales y regulatorios de la IA. En M. Cantero Gamito & L. Bosoer (Eds.), REDemocracIA: Guía para una aproximación democrática a la inteligencia artificial en Iberoamérica (3-51). San Domenico di Fiesole: Instituto Universitario Europeo. Recuperado de: https://cadmus.eui.eu/bitstream/handle/1814/78169/REDemocracIA_Guia_2025.pdf?sequence=4&isAllowed=y.

CLAD (2023). Carta Iberoamericana de Inteligencia Artificial en la Administración Pública. Recuperado de: https://clad.org/declaraciones-consensos/carta-iberoamericana-de-inteligencia-artificial-en-la-administracion-publica/.

Coeckelbergh, M. (2022). The political philosophy of AI. Cambridge: Polity Press.

Collett, C., Neff, G. & Gouvea Gomes, L. (2022). Los efectos de la IA en la vida laboral de las mujeres. UNESCO y OCDE. Recuperado de: https://unesdoc.unesco.org/ark:/48223/pf0000380871.

Council of the European Union (2023). León Declaration on European Neurotechnology: A human-centric and rights-oriented approach 2023. Informal Meeting of Telecommunications Ministers. Recuperado de: https://parleu2023.es/declaration-of-leon-on-parliamentarism.

D'Ignazio, C. & Klein, L. (2023). Feminismo de datos. Cambridge: MIT Press. Recuperado de: https://data-feminism.mitpress.mit.edu/pub/bchsq3az.

Díaz, C. (1997). El ciclo de las políticas públicas locales: Notas para su abordaje y reconstrucción. En J. C. Venecia (Comp.), Políticas públicas y desarrollo local (43-73). FLACSO.

Filgueiras, F. (2023). Desafíos de gobernanza de inteligencia artificial en América Latina: Infraestructura, descolonización y nueva dependencia. Revista del CLAD Reforma y Democracia, 87, 44-70. Recuperado de: https://revista.clad.org/ryd/article/view/desafios-gobernanza-inteligencia-artificial-America-Latina.

Future of Life Institute (2023). Pause Giant AI Experiments: An Open Letter. Recuperado de: https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

García-López, E., Muñoz, J. M. & Andorno, R. (2021). Editorial: Neurorights and mental freedom: Emerging challenges to debates on human dignity and neurotechnologies. Frontiers in Human Neuroscience, 15, 823570. DOI: https://doi.org/10.3389/fnhum.2021.823570.

Gómez Mont, C., May Del Pozo, C., Martínez Pinto, C. & Martín del Campo Alcocer, A. (2024). La inteligencia artificial al servicio del bien social en América Latina y el Caribe: Panorámica regional e instantáneas de doce países. C Minds y Grupo BID. Recuperado de: https://publications.iadb.org/es/la-inteligencia-artificial-al-servicio-del-bien-social-en-america-latina-y-el-caribe-panor%C3%A1mica-regional-e-instant%C3%A1neas-de-doce-paises.

Ienca, M. (2021). On Neurorights. Frontier in Human Neuroscience, 15, 701258. DOI: https://doi.org/10.3389/fnhum.2021.701258.

Ienca, M. & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. Life Sciences, Society and Policy, 13, 5. DOI: https://doi.org/10.1186/s40504-017-0050-1.

Innerarity Grau, D. (2013). Un mundo de todos y de nadie. Buenos Aires: Ediciones Paidós.

Innerarity Grau, D. (2025). Introducción. En M. Cantero Gamito & L. Bosoer (Eds.), REDemocracIA: Guía para una aproximación democrática a la inteligencia artificial en Iberoamérica. San Domenico di Fiesole: Instituto Universitario Europeo. Recuperado de: https://cadmus.eui.eu/bitstream/handle/1814/78169/REDemocracIA_Guia_2025.pdf?sequence=4&isAllowed=y.

Lamm, E., González Alarcón, N. & Flores Urich Sass, A. (2023). Construyendo una inteligencia artificial ética desde el diseño: La perspectiva de la UNESCO. En RICYT (Ed.), El Estado de la Ciencia 2023 (61-72). Buenos Aires: OEI & UNESCO. Recuperado de: https://oei.int/oficinas/argentina/noticias/dossier-sobre-inteligencia-artificial-en-el-estado-de-la-ciencia-2023/.

López Baroni, M. (2019). Las narrativas de la inteligencia artificial. Revista de Bioética y Derecho, 46, 5-28. DOI: https://doi.org/10.1344/rbd2019.0.27280.

Majone, G. (1997). Evidencia, Argumentación y Persuasión en la Formulación de Políticas. México: Fondo de Cultura Económica.

Mantegna, M. (2022). No soy robot: Construyendo un marco ético accionable para analizar las dimensiones de impacto de la inteligencia artificial. Buenos Aires: Universidad de San Andrés.

Mantegna, M. (2024). Datos neuronales a juicio: La Corte Suprema de Chile aborda el primer caso mundial de neuroderechos. En M. Moisés Sánchez, C. Colombara & N. Monti (Eds.), En defensa de los neuroderechos (147-156). Santiago de Chile: KAMANAU. Recuperado de: https://defensaneuroderechos.org.



Monti, N. (2025). A propósito del proceso previo ante la OEA de la Declaración de Principios Interamericanos en materia de neurociencias, neurotecnologías y derechos humanos. En M. Cantero Gamito & L. Bosoer (Eds.), REDemocracIA: Guía para una aproximación democrática a la inteligencia artificial en Iberoamérica (93-97). San Domenico di Fiesole: Instituto Universitario Europeo. Recuperado de: https://cadmus.eui.eu/bitstream/handle/1814/78169/REDemocracIA_Guia_2025.pdf?sequence=4&isAllowed=y.

Oszlak, O. & O´Donnell, G. (1976). Estado y políticas estatales en América Latina: Hacia una estrategia de investigación. Buenos Aires: CEDES.

Peller, J. (2023). Capítulo 4: IA y riesgo existencial en OK Pandora. Buenos Aires: El Gato y la Caja.

Pedace, K., Balmaceda Huarte, T. & Schleider, T. (2023). What Artificial Intelligence is hiding Microsoft and vulnerable girls in northern Argentina. En State of Power 2023: Digital Power (76-84). Amsterdam: Transnational Institute. Recuperado de: https://www.tni.org/en/article/what-artificial-intelligence-is-hiding.

Rikap, C. (2023). Inteligencia artificial: Reemplazo, hibridación… ¿progreso? Nueva Sociedad, 307, 67-81. Recuperado de: https://nuso.org/articulo/307-inteligencia-artificial-reemplazo-hibridacion-progreso/.

Rikap, C. (2025). DeepSeek, China y la IA: un desafío geopolítico. Opinión. Nueva Sociedad, febrero. Recuperado de: https://nuso.org/articulo/deepseek-china-y-la-ia-un-desafio-geopolitico/.

Sadin, E. (2024). La vida espectral: Pensar la era del metaverso y las inteligencias artificiales generativas. Buenos Aires: Caja Negra.

Siverino, P. (2025). Neurotecnologías y derechos humanos: El caso chileno a la luz de los aportes de instrumentos internacionales. En M. Cantero Gamito & L. Bosoer (Eds.), REDemocracIA: Guía para una aproximación democrática a la inteligencia artificial en Iberoamérica (98-102). San Domenico di Fiesole: Instituto Universitario Europeo. Recuperado de: https://cadmus.eui.eu/bitstream/handle/1814/78169/REDemocracIA_Guia_2025.pdf?sequence=4&isAllowed=y.

UNESCO (2021). Report of the International Bioethics Committee of UNESCO (IBC) on the ethical issues of neurotechnology. Recuperado de: https://unesdoc.unesco.org/ark:/48223/pf000037872.



UNESCO (2023). Conferencia Internacional sobre la Ética de la Neurotecnología. Recuperado de: https://www.unesco.org/es/articles/etica-de-la-neurotecnologia.

# Sesgos algorítmicos en los sistemas de inteligencia artificial implementados en administración pública. Un análisis de casos sobre modelos de predicción de riesgo

# Vieses algorítmicos em sistemas de inteligência artificial implementados na administração pública. Uma análise de caso sobre modelos de previsão de risco

# *Algorithmic Biases in Artificial Intelligence Systems Implemented in Public Administration. A Case Analysis of Risk Prediction Models*

**Eugenio Arguelles Toache** *

El uso de la inteligencia artificial (IA) en la administración pública ha crecido significativamente en la última década debido a los múltiples beneficios de esta tecnología en las funciones de gobierno. Sin embargo, esto también ha dejado en evidencia una serie de efectos adversos como la presencia de sesgos algorítmicos que derivan en decisiones que pueden ser discriminatorias e injustas para ciertos individuos y grupos sociales. El objetivo de esta investigación es identificar y analizar las causas de los sesgos algorítmicos presentes en los sistemas de IA de la administración pública, así como sus consecuencias éticas, políticas y sociales. Para ello se analizan cuatro casos representativos a nivel internacional de modelos de predicción de riesgo basados en IA que son utilizados en diferentes ámbitos de la administración pública. El resultado de este análisis muestra que los sesgos algorítmicos se producen en la etapa de diseño y de entrenamiento de los algoritmos, generando afectaciones importantes en el acceso a derechos fundamentales, servicios públicos y programas sociales. La sociedad civil juega un papel fundamental al momento de evidenciar los sesgos algorítmicos y sus consecuencias, así como para contribuir a generar estrategias para la implementación ética de esta tecnología.

**Palabras clave**: modelos de predicción de riesgos; sesgos algorítmicos; discriminación algorítmica; justicia algorítmica; gobierno electrónico

* Becario posdoctoral SECIHTI adscrito al Instituto de Investigaciones Sociales de la Universidad Nacional Autónoma de México (IIS-UNAM). Doctor en ciencias sociales en el área de economía y gestión de la innovación por la Universidad Autónoma Metropolitana (UAM) de México. Correo electrónico: eugenio.toache@gmail.com. ORCID: https://orcid.org/0000-0002-0121-1681.

A utilização da inteligência artificial (IA) na administração pública tem crescido significativamente na última década devido aos múltiplos benefícios desta tecnologia nas funções governamentais. No entanto, isto também revelou uma série de efeitos adversos, como a presença de enviesamentos algorítmicos que geram resultados e decisões que podem ser discriminatórias e injustas para alguns indivíduos ou grupos sociais. O objetivo desta investigação é identificar e analisar as causas dos enviesamentos algorítmicos presentes nos sistemas de IA da administração pública, bem como as suas consequências éticas, políticas e sociais. Para tal, são analisados quatro casos internacionalmente representativos de modelos de previsão de risco baseados em IA, utilizados em diferentes áreas da administração pública. Os resultados desta análise mostram que os enviesamentos algorítmicos ocorrem nas etapas de conceção e treino dos algoritmos, gerando impactos significativos no acesso a direitos fundamentais, serviços públicos e programas sociais. A sociedade civil desempenha um papel fundamental na exposição dos enviesamentos algorítmicos e das suas consequências, bem como na contribuição para a geração de estratégias para a implementação ética desta tecnologia.

**Palavras-chave:** modelos de previsão de risco; enviesamentos algorítmicos; discriminação algorítmica; justiça algorítmica; governo eletrônico

*The use of artificial intelligence (AI) in public administration has grown significantly in the last decade due to the multiple benefits of this technology in government functions. However, this has also revealed a series of adverse effects such as the presence of algorithmic biases that generate results and decisions that can be discriminatory and unfair for some individuals or social groups. The objective of this article is to identify and analyze the causes of algorithmic biases present in AI systems in public administration, as well as their ethical, political, and social consequences. To do so, four internationally representative cases of AI-based risk prediction models that are used in different areas of public administration are analyzed. The result of this analysis shows that algorithmic biases occur at the design and training stage of the algorithms, generating significant impacts on access to fundamental rights, public services, and social programs. Civil society plays a fundamental role in exposing algorithmic biases and their consequences, as well as in contributing to the generation of strategies for the ethical implementation of this technology.*

***Keywords**: risk prediction models; algorithmic biases; algorithmic discrimination; algorithmic justice; e-government*

## Introducción

En la última década, se ha evidenciado un crecimiento significativo en el uso de la inteligencia artificial (IA) por parte de la administración pública. Este crecimiento se aceleró considerablemente tras la pandemia de COVID-19, marcando un punto de inflexión en la adopción de tecnologías digitales para la gestión pública. La IA está transformando la forma en que los gobiernos operan, enfocándose en la optimización de procesos administrativos internos, la formulación más precisa de políticas públicas, la toma de decisiones basada en datos, la mejora en la prestación de servicios públicos y la creación de canales de comunicación más eficientes con la ciudadanía. Las principales áreas en las que la administración pública ha implementado IA incluyen la salud pública, la seguridad y el orden público, el transporte público y la gestión de la movilidad, la prevención y la gestión de desastres naturales, la asistencia social y la administración tributaria.

No obstante, el uso de la IA en el ámbito gubernamental ha suscitado un extenso debate en la academia y la sociedad en general sobre sus posibles beneficios y los potenciales efectos adversos. Entre los efectos adversos más discutidos destacan los sesgos algorítmicos, un fenómeno ampliamente documentado en la literatura especializada. Por ejemplo, Valle-Cruz *et al.* (2024) llevaron a cabo una revisión sistemática de investigaciones previas y concluyeron que los sesgos algorítmicos representan uno de los impactos negativos más frecuentes y preocupantes del uso de la IA en la administración pública. Los sesgos algorítmicos son tendencias o desviaciones que influyen en las valoraciones, estimaciones o predicciones generadas por los algoritmos y que afectan de manera desproporcionada a determinados individuos o grupos sociales. La pregunta general que guía esta investigación es: ¿cuáles son las causas y consecuencias de los sesgos algorítmicos en los sistemas de IA utilizados por la administración pública y cuáles son las estrategias más utilizadas para eliminar o mitigar estos sesgos y sus consecuencias?



Los sesgos algorítmicos provienen principalmente de dos fuentes. En primer lugar, los sesgos algorítmicos se pueden producir en el proceso de diseño cuando se toman decisiones que reflejan consciente o inconscientemente los perjuicios de quienes los desarrollan o cuando se toman decisiones erróneas que generan desviaciones en los parámetros y reglas que conforman el algoritmo. En segundo lugar, los sesgos algorítmicos se pueden producir en el proceso de entrenamiento cuando los algoritmos son entrenados con datos históricos que reflejan prejuicios sociales o desigualdades sistémicas, datos que no son representativos o que han sido erróneamente etiquetados. Como resultado, los sistemas algorítmicos no solo replican estas desigualdades y estos perjuicios, sino que, debido a su capacidad de automatizar y escalar procesos, pueden amplificarlas a gran escala.

Los sesgos presentes en los algoritmos que guían decisiones gubernamentales tienen el potencial de provocar efectos adversos significativos, como la perpetuación o amplificación de dinámicas de exclusión y discriminación estructural. En concreto,

los sesgos algorítmicos pueden traducirse en políticas públicas y decisiones administrativas que marginen sistémicamente a ciertos grupos sociales en función de características como su nacionalidad, origen étnico, género, idioma o religión. Estas dinámicas discriminatorias no solo refuerzan desigualdades históricas, sino que también obstaculizan el acceso equitativo a derechos fundamentales, servicios públicos esenciales y programas sociales diseñados para mejorar la calidad de vida de la población. Ante esta realidad, el análisis de los sesgos algorítmicos, más que un tema técnico, es un desafío ético y político que exige la responsabilidad de los gobiernos para evitar que la IA refuerce desigualdades. Para ello es esencial establecer marcos regulatorios sólidos, realizar auditorías independientes, fomentar la transparencia y garantizar la participación de diversos actores sociales. Abordar estos sesgos no solo es una cuestión de justicia social, sino también de legitimidad institucional, asegurando que la IA contribuya a la equidad y el bienestar colectivo en lugar de perpetuar la discriminación.

El objetivo de este artículo de investigación es examinar una selección de casos representativos a nivel internacional para identificar y analizar los sesgos algorítmicos presentes en los modelos de predicción de riesgos basados en IA utilizados en la administración pública, enfocando el análisis en tres aspectos principales: i) las causas de estos sesgos vinculadas tanto al diseño de los algoritmos como a los datos empleados en su entrenamiento; ii) sus consecuencias en términos de sus afectaciones sobre el acceso a derechos fundamentales, servicios públicos y programas sociales; y iii) las principales estrategias utilizadas para abordar y mitigar dichos sesgos, destacando las mejores prácticas y lecciones aprendidas de cada caso estudiado.



Los cuatro casos que se analizarán son: COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), Algoritmo AMS (*Arbeits Markt Service*) y SyRI (*System Risk Indication*) y la Plataforma Tecnológica de Intervención Social. La documentación de cada caso se realizará a través del análisis de diversas fuentes, incluyendo artículos científicos, reportajes periodísticos, documentos institucionales y contenidos disponibles en sitios web. Previo a eso, el siguiente apartado ofrece una revisión y discusión de la literatura científica sobre los sesgos algorítmicos.

## 1. Sesgos algorítmicos: definición, causas y consecuencias

De acuerdo con Simon *et al*. (2020), los sesgos algorítmicos se refieren a errores recurrentes en sistemas computacionales que, de manera injusta, discriminan a ciertos individuos o grupos sobre otros al asignarles resultados desfavorables basados en criterios inapropiados o irracionales y perpetuando las desigualdades presentes en la sociedad. Pasquinelli *et al.* (2022) complementan esta idea y observan que los sesgos algorítmicos son, en esencia, una amplificación de prejuicios preexistentes en los datos y en la incorporación inadvertida de prejuicios humanos en la estructura y las reglas de los algoritmos, lo cual ocurre debido a errores computacionales, una comprensión inadecuada de la información o deficiencias en las técnicas de aproximación utilizadas

en el aprendizaje automático. En síntesis, los sesgos algorítmicos son manifestaciones de patrones, tendencias y errores que subyacen en los algoritmos y que conducen a estimaciones, predicciones y valoraciones desiguales que favorecen o perjudican a ciertos individuos o grupos sociales, generando resultados que, en última instancia, son discriminatorios, excluyentes e injustos.

Los sesgos algorítmicos, según CAF (2021) y Martín (2022), tienen su origen esencialmente en dos fuentes principales: el diseño de los algoritmos y los datos utilizados para entrenarlos. Los sesgos en el proceso de diseño de los algoritmos se deben a que estos son diseñados por seres humanos que toman decisiones sobre las reglas y los parámetros que definen la estructura de los algoritmos, de modo que los sesgos algorítmicos surgen de estas decisiones. En primer lugar, los seres humanos pueden, de manera involuntaria, incorporar sus propias experiencias, valores, perspectivas del mundo y prejuicios durante el proceso de diseño de algoritmos, por lo que dichas influencias se manifestarán en la configuración de parámetros, la definición de reglas y la selección de criterios de evaluación, lo que da lugar a modelos con tendencias y desviaciones inherentes (Martín, 2022). En segundo lugar, los diseñadores de los algoritmos pueden tomar, de manera involuntaria, decisiones erróneas sobre las variables, métricas y parámetros que componen a los algoritmos, generando igualmente tendencias y desviaciones en los modelos (CAF, 2021). En tercer lugar, aunque la mayoría de estos sesgos se introduce de manera involuntaria, también pueden ser añadidos intencionalmente para manipular o engañar a ciertos grupos de individuos, lo que plantea implicaciones éticas graves (Sued, 2022). De esta manera, los algoritmos dejan de ser neutrales y reproducen las subjetividades de sus creadores, influyendo en cómo se interpretan y procesan las necesidades y características de diferentes grupos sociales



Por otro lado, los sesgos algorítmicos también emergen de las características de los datos utilizados para entrenar los modelos algorítmicos y sus procesos de aprendizaje automático. En primer lugar, estos datos pueden contener sesgos históricos, es decir, prejuicios acumulados a lo largo del tiempo como resultado de decisiones humanas y prácticas pasadas, estos sesgos se perpetúan y amplifican cuando los algoritmos los replican sin cuestionarlos (Baker & Hawn, 2022; Koniakou, 2023). En segundo lugar, los datos utilizados para entrenar a los algoritmos pueden ser incompletos o no representativos de ciertas poblaciones, lo que lleva a que el algoritmo produzca generalizaciones erróneas y resultados sesgados (Kordzadeh & Ghasemaghaei, 2022). Finalmente, incluso si los datos son completos y representativos y no contienen sesgos históricos, las decisiones humanas tomadas durante el proceso de clasificación y etiquetado, como la selección, eliminación y priorización de algunas variables y datos, pueden introducir sesgos inadvertidos (CAF, 2021)

En resumen, los sesgos algorítmicos son una extensión de los prejuicios humanos y pueden surgir en distintas etapas del ciclo de vida de un algoritmo. Aparecen en el diseño, cuando los desarrolladores incorporan reglas sesgadas, en el entrenamiento, al usar datos incompletos o con prejuicios históricos, y en la evaluación, al validar los

resultados con métricas sesgadas (CAF, 2021). Aunque en términos generales los sesgos algorítmicos son el reflejo de los sesgos humanos, sus consecuencias pueden ser significativamente más graves y de mayor alcance. Esto se debe a que la IA se encuentra en un proceso de adopción masiva y crecimiento exponencial tanto en la industria como en la administración pública; por lo tanto, existe el riesgo de amplificar y perpetuar estos sesgos a una escala sin precedentes. Es decir, el uso intensivo y generalizado de la IA permite que los sesgos algorítmicos no solo se proyecten sobre una gran cantidad de personas, sino que también extiendan sus efectos adversos en múltiples contextos y a lo largo del tiempo (Martín, 2022).

Los sesgos algorítmicos generan una representación desproporcionada de ciertos grupos sociales frente a otros, lo que puede ocurrir de manera sistémica al basarse en características como raza, género, nacionalidad o clase social. Esto implica que los algoritmos pierden la neutralidad y objetividad esperadas, ya que introducen desviaciones y tendencias que resultan en un trato desigual, así como en predicciones o valoraciones que no son equitativas entre diferentes grupos identitarios (Kordzadeh & Ghasemaghaei, 2022). Una vez que los sistemas de IA comienzan a operar, los sesgos integrados en ellos se replican constantemente, produciendo resultados excluyentes, discriminatorios e inequitativos. Estos resultados, a su vez, pueden ser utilizados como datos de entrada para entrenar nuevamente los algoritmos, creando un círculo vicioso de retroalimentación sesgada. Este proceso amplifica y perpetúa las desigualdades sociales, ya que cada interacción refuerza los prejuicios integrados en el sistema (Martín, 2022; Koniakou, 2023).



A diferencia de los sesgos humanos, que están limitados por las interacciones individuales, los sesgos algorítmicos tienen un alcance exponencial. La capacidad de los algoritmos para procesar y tomar decisiones de manera autónoma, en combinación con su uso extendido en sectores como la salud, la educación, las finanzas y la justicia, multiplica el impacto de estos sesgos, afectando a millones de personas en todo el mundo. Además, al no ser detectados o corregidos adecuadamente, estos sesgos se convierten en estructuras rígidas que perpetúan desigualdades históricas bajo una apariencia de objetividad tecnológica.

El riesgo asociado a los sesgos algorítmicos en los sistemas de IA empleados en la administración pública es especialmente preocupante, debido a la magnitud y el impacto potencial de las decisiones que estos sistemas generan. A diferencia de otros contextos, en el ámbito de la administración pública las decisiones basadas en algoritmos tienen el potencial de afectar directamente a un volumen significativo de la población durante periodos prolongados de tiempo. Esto implica que los sesgos presentes no solamente se reproducen y amplifican en una escala mayor, sino que también tienden a perpetuarse, reforzando desigualdades existentes y creando nuevas barreras estructurales (Martín, 2022).

Adicionalmente, el uso de sistemas algorítmicos en la administración pública tiene una particular trascendencia debido a las áreas estratégicas y prioritarias en las que suelen

aplicarse. Estos sistemas influyen en aspectos fundamentales para el bienestar social, como el acceso a derechos humanos, la distribución de recursos en programas sociales y la prestación de servicios públicos esenciales, como salud, educación y justicia (CAF, 2021; Koniakou, 2023). De esta manera, cuando los sesgos algorítmicos se filtran en la toma de decisiones gubernamentales, se convierten en un problema no solo técnico, sino también profundamente ético, político y de justicia social. Las decisiones injustas basadas en algoritmos pueden consolidar dinámicas de exclusión, discriminación y marginación, especialmente hacia grupos históricamente desfavorecidos. Estas dinámicas, además, suelen ser menos visibles y más difíciles de cuestionar, debido al aparente carácter objetivo y neutral de las tecnologías algorítmicas.

En esta investigación se analizan los modelos de predicción de riesgos basados en IA, los cuales analizan grandes volúmenes de datos para identificar patrones y estimar la probabilidad de que ocurran ciertos eventos adversos. En la administración pública, estos modelos se utilizan para optimizar la toma de decisiones en áreas como seguridad, salud, bienestar social y justicia. Por ejemplo, pueden evaluar el riesgo de reincidencia delictiva, predecir la vulnerabilidad de ciertos grupos o detectar posibles fraudes en programas sociales.

La hipótesis de esta investigación es que los sesgos algorítmicos en los modelos de predicción de riesgo utilizados en la administración pública se originan tanto en el diseño del algoritmo, debido a la selección sesgada de variables predictivas, como en su entrenamiento, al emplear datos que reflejan desigualdades estructurales. Estos sesgos pueden llevar a la sobreestimación del riesgo en grupos vulnerables, reforzando patrones de discriminación y exclusión, lo que impacta el acceso equitativo a derechos, programas sociales y servicios públicos. En el siguiente apartado se describe la metodología utilizada para realizar el análisis de las causas y consecuencias de los sesgos algorítmicos presentes en los sistemas de IA utilizados en la administración pública.



## 2. Metodología

Para la selección de los casos de estudio, se empleó la base de datos del Observatorio de Algoritmos con Impacto Social (OASI, por sus siglas en inglés) de Éticas Foundation, la cual recopila y analiza 152 casos a nivel global en los que el uso de algoritmos de IA ha generado consecuencias sociales negativas no previstas. El propósito de esta base de datos es visibilizar la problemática y fomentar la conciencia sobre la necesidad de auditar y regular el uso de algoritmos en ámbitos que impactan directamente a la sociedad. De los 152 casos documentados, 96 corresponden a algoritmos implementados en la administración pública y, tras un primer análisis exploratorio, se identificaron 24 casos específicamente relacionados con modelos algorítmicos diseñados para la predicción de riesgos. Posteriormente, estos 24 casos fueron examinados puntualmente, priorizando aquellos que contaban con suficiente información detallada para un análisis riguroso y priorizando la diversidad de los casos

en cuanto a su ámbito de implementación y los países de origen. Como resultado de este proceso de selección, se eligieron cuatro casos para su estudio exhaustivo, permitiendo así una evaluación más detallada de las causas de los sesgos algorítmicos y de sus impactos políticos y sociales.

Los casos seleccionados son los siguientes: i) Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), algoritmo utilizado en el sistema judicial estatal de los Estados Unidos para predecir el riesgo de reincidencia delictiva; ii) System Risk Indication (SyRI), algoritmo implementado en Países Bajos para predecir el riesgo de fraude en solicitudes de subsidios y apoyos sociales; iii) Allegheny Family Screening Tool (AFST), algoritmo implementado en el condado de Allegheny en Pensilvania, Estados Unidos, para predecir el riesgo de negligencia y abuso infantil; y iv) Plataforma Tecnológica de Intervención Social, algoritmo implementado en la provincia de Salta, Argentina, para predecir el riesgo de deserción escolar y de embarazo en adolescentes.

Para recopilar información sobre los casos de estudio, se empleó un enfoque basado en el análisis documental y la revisión de sitios web. La documentación se sustentó en tres tipos de fuentes principales: artículos científicos, documentos y reportes institucionales, y fuentes hemerográficas. La obtención de estos documentos se llevó a cabo mediante la búsqueda del nombre exacto de cada caso en bases de datos académicas como Scopus y Google Scholar, así como en el motor de búsqueda de Google. Con el objetivo de garantizar la rigurosidad del análisis, se estableció un criterio de prioridad en el uso de las fuentes, dando preferencia a los artículos científicos, seguidos de los reportes institucionales y, en última instancia, de las fuentes hemerográficas. En cuanto al análisis de sitios web, se realizó una búsqueda de las páginas oficiales de los proyectos estudiados o de las organizaciones responsables de su desarrollo e implementación, con el propósito de obtener información directa y verificada sobre cada caso. Adicionalmente, para fortalecer el análisis de los casos estudiados, se incluyen ejemplos de casos similares implementados en otros países, lo que permiten contextualizar y enriquecer los hallazgos obtenidos.



## 3. Análisis de casos

### 3.1. Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)

El sistema COMPAS es uno de los modelos de predicción de riesgo basados en IA más conocidos y fue desarrollado por la empresa Northpointe en el año 2000. Este sistema es ampliamente utilizado en el sistema judicial de los Estados Unidos para predecir el riesgo de reincidencia delictiva y apoyar el proceso de toma de decisiones de los jueces en torno a la concesión de libertad bajo fianza antes del juicio o de la libertad condicional de los condenados, una vez que han cumplido una parte de la sentencia. Mediante un algoritmo de IA, COMPAS analiza hasta 12 variables personales, como

la edad, los antecedentes penales y los problemas de adicciones o educación, para asignar a cada detenido o condenado una puntuación de riesgo del 1 al 10. Estas puntuaciones se agrupan en tres niveles: bajo nivel de reincidencia (1-4), medio nivel de reincidencia (5-7) y alto nivel de reincidencia (8-10). Aquellos clasificados como de alto riesgo tienen muy pocas posibilidades de obtener libertad bajo fianza o libertad condicional; aquellos clasificados como de riesgo medio pueden acceder a este beneficio con condiciones de vigilancia estricta; mientras que aquellos clasificados como de bajo riesgo suelen recibir este beneficio con menores restricciones (Blomberg *et al*., 2010).

El sistema COMPAS se ha convertido en un caso emblemático y ampliamente citado de sesgos algorítmicos en la administración pública. En 2016, gracias a una investigación periodística de ProPublica, se descubrió la existencia de un sesgo algorítmico racial en este sistema. La investigación de Angwin *et al*. (2016) analizó 11.757 casos que fueron evaluados mediante COMPAS y encontró que este sistema clasificaba de manera desproporcionada a personas de piel negra con un riesgo más alto de reincidencia en comparación con personas de piel blanca; en concreto, las personas de piel negra fueron clasificadas erróneamente como de alto riesgo casi el doble de veces que las personas de piel blanca que no reincidieron (45% frente a 23%). Incluso ajustando por factores como delitos previos, edad y género, los acusados de piel negra tenían un 45% más de probabilidades de recibir puntuaciones más altas de riesgo general y un 77% más de probabilidades de alto riesgo de reincidencia violenta.



En menos de 18 meses, más de 200 artículos académicos citaron el artículo de ProPublica y emplearon métodos alternativos para reproducir el análisis llegando a resultados similares y a la misma conclusión: el sistema COMPAS presenta un sesgo algorítmico racial (Washington, 2018). Por ejemplo, la investigación de Hamilton (2019) utilizó la misma base de datos que la investigación de ProPublica y encontró que COMPAS también presenta un sesgo algorítmico hacia las personas hispanas al asignarles clasificaciones que sobrestiman su riesgo de reincidencia.

Autores como Fuchs (2018), Avella *et al*. (2022) y Marinucci *et al.* (2023) señalan que los sesgos algorítmicos presentes en el sistema COMPAS tienen su origen en el uso de datos históricos marcados por prejuicios, desigualdades estructurales y discriminación sistémica en el sistema legal, lo cual se trasfiere al algoritmo mediante el proceso de entrenamiento. Estos datos incluyen patrones desproporcionados de informes de delitos provenientes de vecindarios predominantemente habitados por personas afroamericanas e hispanas, reflejando las desigualdades y los prejuicios sociales que afectan a estas comunidades. Según Avella *et al.* (2022), las comunidades históricamente sometidas a una vigilancia y persecución desproporcionadas por parte de las fuerzas del orden, en particular las de bajos ingresos y las minorías raciales, enfrentan ahora un nuevo nivel de discriminación, ya que el algoritmo tiende a calificarlas con puntuaciones más altas de reincidencia. Este fenómeno perpetúa y amplifica las desigualdades existentes, transformando las disparidades humanas en decisiones algorítmicas con consecuencias significativas para la equidad y la justicia.

Adicionalmente, estudios recientes han identificado que el sistema COMPAS también exhibe un sesgo algorítmico relacionado con el género, el cual tiende a sobreestimar el riesgo de reincidencia en mujeres. Este sesgo se debe a un error en el proceso de diseño del algoritmo, ya que fue diseñado para ser neutral en términos de género y evitar prejuicios; sin embargo, termina generando el efecto contrario. Al no considerar el género como una variable predictiva, evalúa el riesgo de reincidencia de manera uniforme para hombres y mujeres, pese a que las mujeres presentan tasas significativamente menores de reincidencia. Este sesgo incrementa la probabilidad de que sean tratadas de manera injusta en el ámbito de la justicia penal, al enfrentarse a decisiones como la negación de la libertad bajo fianza o condicional, o sean objeto de niveles de supervisión excesivos en caso de obtener dichos beneficios, afectando con ello sus derechos fundamentales (Skeem *et al*., 2016; Hamilton, 2018). Para mitigar este sesgo de género en el sistema COMPAS, sus desarrolladores han implementado recientemente normas específicas para hombres y mujeres, ajustando la clasificación de las puntuaciones según el género. No obstante, la adopción de estas normas depende de los funcionarios judiciales de cada Estado o condado, quienes en su mayoría han rechazado su aplicación, argumentando que la neutralidad de género debe prevalecer al tratarse de un grupo protegido que no debería influir en las decisiones de justicia penal (Hamilton, 2018).

Como consecuencia de los sesgos algorítmicos presentes en el sistema COMPAS, no solamente se reproducen las desigualdades estructurales presentes en la sociedad, sino que, al ser utilizado para tomar decisiones judiciales clave, como otorgar libertad provisional o condicional, las perpetúa y amplifica. Esto afecta la equidad en el acceso a beneficios judiciales para las personas de piel negra, otras minorías y mujeres, resultando especialmente afectadas las mujeres afroamericanas, erosionando la confianza en la imparcialidad del sistema de justicia. Este ciclo de retroalimentación entre datos sesgados y decisiones basadas en ellos agrava las barreras estructurales existentes y contribuye a mantener dinámicas de discriminación institucionalizada (CAF, 2021).



A pesar de las evidencias que señalan la existencia de sesgos algorítmicos en el sistema COMPAS, la empresa Northpointe ha rechazado estas afirmaciones, argumentando que su herramienta es precisa para predecir la reincidencia general, los actos violentos y la incomparecencia ante el tribunal. Además, el algoritmo de COMPAS está protegido por leyes de secreto comercial, lo que impide que se conozca su funcionamiento exacto y dificulta el análisis y la deliberación en contextos legales. Esta falta de transparencia, sumada a la falta de un examen exhaustivo de las implicaciones constitucionales por parte de la Corte Suprema de los Estados Unidos, ha dejado el tema sin una resolución definitiva, ya que las discusiones en tribunales estatales y federales inferiores no han logrado abordar plenamente la problemática.

Alrededor del mundo existen casos similares al sistema COMPAS que igualmente han sido objeto de críticas por la presencia de sesgos algorítmicos. Por ejemplo, en Cataluña, España, se utiliza la herramienta RisCanvi (acrónimo de las palabras

"riesgo" y "cambio" en catalán) y -aunque la única investigación independiente que existe demuestra que esta herramienta no presenta sesgos algorítmicos de género, raza, nacionalidad o edad- Aróstegui (2023) indica que el uso de datos de entrenamiento desactualizados y no representativos puede generar predicciones de riesgo sesgadas sobre los grupos menos representados en las prisiones de Cataluña, como las mujeres y los extranjeros.

### 3.2. System Risk Indication (SyRI)

En 2012, el gobierno de Países Bajos diseñó e implementó el modelo de predicción de riesgo conocido como SyRI con el propósito de elaborar indicadores de riesgo para detectar posibles casos de fraude en solicitudes de subsidios y apoyos sociales. Este sistema se basaba en un algoritmo predictivo de IA que procesaba una amplia cantidad de datos personales de los ciudadanos, abarcando aspectos como empleo, sanciones administrativas, información fiscal, propiedades, beneficios sociales, antecedentes educativos, integración cívica, cumplimiento normativo, deudas, pensiones y permisos (Berning, 2023). Posteriormente, el algoritmo comparaba estos datos con perfiles de riesgo construidos a partir de información de personas previamente identificadas como defraudadoras; al identificar similitudes generaba una lista de individuos o entidades con alta probabilidad de incurrir en fraude relacionado con las prestaciones sociales ofrecidas por el gobierno (R3D, 2020). Finalmente, este análisis daba lugar a la elaboración de informes de riesgo que servían para investigar a posibles defraudadores de manera más detallada, lo cual podía derivar en su exclusión o suspensión de los programas sociales, o en acusaciones de fraude (Lazcoz & Castillo, 2020).



En octubre de 2018, un relator especial de la ONU sobre pobreza extrema y derechos humanos remitió un informe al Tribunal de La Haya en el que señalaba serias deficiencias en el funcionamiento del algoritmo SyRI, principalmente la presencia de sesgos algorítmicos que generaban un efecto discriminatorio y estigmatizador hacia los ciudadanos pertenecientes a determinados grupos étnicos y con un menor nivel de ingresos (R3D, 2020; Lazcoz & Castillo, 2020). En posteriores investigaciones se descubrió que los sesgos étnicos se originaron debido a que el algoritmo fue diseñado para clasificar las nacionalidades diferentes a la holandesa como un indicador de riesgo, reflejando una discriminación implícita en su programación (Peeters & Widlak, 2023). Este sesgo fue exacerbado por el uso de datos históricos para entrenar el algoritmo, datos que ya contenían prejuicios estructurales y actitudes discriminatorias contra minorías étnicas, principalmente aquellas con raíces en las antiguas colonias holandesas de Surinam o en las islas caribeñas holandesas, así como grupos con herencia en Turquía y Marruecos (Arts & Van den Berg, 2024). Igualmente, los sesgos socioeconómicos se presentaron debido a que los datos utilizados para entrenar al algoritmo reflejaban discriminación hacia personas que habitaban zonas caracterizadas por sus bajos ingresos (Berning, 2023). Como resultado, SyRI perpetuó y amplificó estas desigualdades estructurales al generar clasificaciones que asociaban desproporcionadamente un mayor riesgo a las personas de nacionalidades

diferentes, los grupos étnicos minoritarios y las personas con bajos ingresos. La falta de auditorías externas e independientes para evaluar los efectos adversos de este sistema permitió que los resultados discriminatorios se perpetuaran y retroalimentaran a lo largo del tiempo.

Los sesgos algorítmicos presentes en el sistema SyRI y sus resultados discriminatorios tuvieron un severo impacto en la vida de miles de personas. De acuerdo con Peeters y Widlak (2023), entre 2012 y 2019, más de 30.000 ciudadanos fueron injustamente acusados de cometer fraude, lo que desencadenó en la pérdida de beneficios sociales fundamentales, la exclusión de programas de asistencia gubernamental y una grave situación de endeudamiento, ya que muchos de ellos fueron obligados a reembolsar todos los beneficios económicos que habían recibido hasta el momento. Para las personas afectadas, en su mayoría provenientes de comunidades migrantes o de bajos ingresos, las consecuencias de estas acciones resultaron devastadoras y, en muchos casos, irreparables. Estas medidas no solo las empujaron a situaciones de pobreza extrema y, en algunos casos, a la bancarrota, sino que también afectaron profundamente múltiples aspectos de sus vidas; por ejemplo, muchas personas informaron haber desarrollado problemas de salud mental, como depresión y estrés crónico, y otras sufrieron el estigma social de ser consideradas defraudadoras afectando permanentemente sus relaciones personales y sociales, mientras que las madres y padres de más de 2000 infantes perdieron su custodia como consecuencia de estas decisiones discriminatorias, lo cual no solo desintegró familias, sino que también generó un trauma intergeneracional que amplificó la vulnerabilidad de los menores y limitó severamente sus oportunidades de un futuro estable (Arts & Van den Berg, 2024).



En 2020, el Tribunal de Distrito de La Haya declaró ilegal el uso del algoritmo SyRI, sentando un precedente significativo en Europa, ya que se trató de la primera sentencia para declarar ilegal un algoritmo en dicho continente. El fallo señaló que este sistema violaba derechos fundamentales, como la privacidad y la protección contra la discriminación, aunado a la falta de transparencia del algoritmo y su incapacidad para evitar sesgos. El caso SyRI desencadenó una crisis política de gran magnitud que provocó la renuncia del tercer gabinete del gobierno del primer ministro Mark Rutte en 2021, mientras que en 2022 el gobierno holandés reconoció formalmente que este sistema y las decisiones derivadas fueron un ejemplo de racismo institucional. A pesar de esto, las consecuencias para las familias afectadas persisten, y muchas de ellas continúan enfrentando largas batallas legales en busca de justicia y de las indemnizaciones correspondientes (Arts & Van den Berg, 2024).

En 2016 se presentó un caso similar a SyRI en Australia cuando se implementó el sistema RoboDebt, un algoritmo de IA diseñado para detectar posibles pagos excesivos en asistencia social mediante la comparación de los ingresos reportados con los subsidios recibidos, si detectaba alguna discrepancia se emitía automáticamente avisos de cobro de deudas a los ciudadanos. Sin embargo, el sistema presentaba un sesgo socioeconómico que afectó principalmente a los estratos socioeconómicos más

bajos, generando errores en el cálculo de sus ingresos y provocando interrupciones en los pagos de asistencia y graves problemas de endeudamiento para miles de ciudadanos, ya que se emitieron indebidamente 470.000 deudas (Rinta-Kahila *et al*., 2022). Tras una investigación y críticas generalizadas, el gobierno australiano eliminó el sistema, enfrentó una demanda colectiva y reembolsó alrededor de 1200 millones de dólares a los afectados. Igualmente, el Departamento de Trabajo y Pensiones (DWP, por sus siglas en inglés) del Reino Unido, implementó en 2016 un algoritmo de predicción de riesgos para detectar posibles fraudes en los beneficios sociales. Sin embargo, en 2021 se reportó que dicho algoritmo presentaba un sesgo hacia las personas con discapacidad, clasificándolas con un mayor riesgo de cometer fraude y resultando en investigaciones intrusivas y humillantes para dichas personas. Esto se tradujo en acciones legales por parte de los afectados para que se diera a conocer el funcionamiento del algoritmo (Savage, 2021).

### 3.3. Allegheny Family Screening Tool (AFST)

En 2016, el Departamento de Servicios Humanos del condado de Allegheny, en Pensilvania, Estados Unidos, implementó el modelo predictivo de riesgo conocido como AFST para predecir e identificar posibles casos de negligencia y abuso infantil con el objetivo de que los trabajadores sociales tomen decisiones más informadas y realicen las intervenciones necesarias para garantizar la seguridad y el bienestar de los infantes. Esta herramienta utiliza un algoritmo de IA para procesar y analizar datos provenientes de diferentes fuentes (Chouldechova *et al*., 2018; Blanchard, 2022): i) llamadas telefónicas relacionadas con posibles casos de negligencia y abuso infantil realizadas por miembros de la comunidad al Departamento de Servicios Humanos; ii) datos históricos de informes que contienen denuncias de maltrato previas o denuncias por abuso de drogas o alcohol; y iii) registros administrativos relevantes como ser beneficiario de algún programa social, recibir tratamiento de salud mental, estar en un programa de libertad condicional o habitar en un vecindario de bajos ingresos. A partir de estas variables predictivas, el algoritmo genera una puntuación de riesgo para cada caso, en una escala del 1 al 20, donde 20 representa el nivel más alto de riesgo. Esta puntuación tiene como objetivo priorizar las acciones de los trabajadores sociales en función del riesgo calculado; por ejemplo, los casos con puntaciones de riesgo bajo generalmente son descartados o archivados, y los casos con puntuaciones de riesgo medio no requieren una intervención inmediata, por lo que los trabajadores sociales establecen un monitoreo continuo o inician una investigación más exhaustiva, mientras que los casos con puntuaciones de riesgo alto requieren una intervención inmediata por parte de los trabajadores sociales: realizar visitas al hogar, aplicación de entrevistas, asesoramiento psicológico, incorporación a programas de intervención familiar o –en el caso más grave– separación de niños de sus familias de origen y posterior asignación a una organización de acogida (Gerchick *et al*., 2023; Rittenhouse *et al*., 2023).



Poco después de su implementación, la herramienta AFST comenzó a recibir críticas de diversos investigadores debido a la presencia de sesgos algorítmicos, particularmente raciales y socioeconómicos. Por ejemplo, las investigaciones de

Eubanks (2017) y Chouldechova *et al.* (2018) revelaron que la herramienta asigna puntuaciones de riesgo desproporcionadamente altas a familias de color y en situación de pobreza. Los sesgos socioeconómicos surgen debido a que el algoritmo está diseñado para utilizar variables predictivas vinculadas directamente con la pobreza, como el uso de beneficios gubernamentales, programas sociales, servicios de salud pública, no tener suficiente comida, tener una vivienda inadecuada o dejar a un niño solo mientras se trabaja (Eubanks, 2017). Aunque estas variables buscan identificar posibles factores asociados con la negligencia o el abuso infantil, terminan reflejando y perpetuando desigualdades estructurales. Adicionalmente, el uso de datos históricamente sesgados en el entrenamiento del algoritmo agrava esta problemática, ya que el Departamento de Servicios Humanos interviene con mayor frecuencia en comunidades de bajos recursos, por lo que se generan más informes sobre estas poblaciones, los cuales asocian históricamente la pobreza con la negligencia y el abuso infantil (Rittenhouse *et al*., 2023). Todo esto se traduce en que el algoritmo establezca una correlación errónea entre la pobreza y la negligencia infantil, lo que resulta en una mayor probabilidad de que estas familias sean clasificadas como de alto riesgo, independientemente de las condiciones reales del entorno familiar.

Por otro lado, los sesgos raciales se originan por la alta cantidad de familias de color y multirraciales en las bases de datos utilizadas por el Departamento de Servicios Humanos. En primer lugar, los denunciantes de la comunidad llaman a las líneas de abuso y negligencia infantil para reportar frecuentemente casos que involucran a familias negras y multirraciales tres veces y media más veces en comparación con los casos relacionados con familias blancas, generando con ello más informes y sobrerrepresentación (Eubanks, 2017). En segundo lugar, los registros administrativos utilizados para alimentar el algoritmo incluyen más información sobre estos grupos raciales en comparación con otros, ya que es más probable que esos grupos participen en programas gubernamentales, lo que refuerza esta sobrerrepresentación (Chouldechova *et al*., 2018). Este patrón de denuncias y de representatividad en los datos refleja prejuicios raciales sistémicos presentes en la sociedad, los cuales se ven amplificados y perpetuados al integrarse en los datos utilizados por el algoritmo.



Estos sesgos algorítmicos generan que las familias de bajos ingresos y de color tengan una mayor probabilidad de ser clasificadas erróneamente como de alto riesgo, lo que resulta en un elevado número de falsos positivos que desencadenan investigaciones y medidas de intervención desproporcionadas, muchas veces discriminatorias e injustas, que pueden llevar incluso a la pérdida de la custodia de los infantes. Además, las puntuaciones obtenidas a través del AFST, junto con las acciones subsecuentes, pueden perpetuar el impacto negativo en estas familias, agravando los efectos de la discriminación sistémica, ya que no solamente limita las oportunidades de defensa y apelación frente a dichas decisiones, sino que también restringe el acceso a otros beneficios sociales, afectando profundamente sus derechos y su capacidad de superar las condiciones que inicialmente las colocaron en una situación vulnerable (Gerchick *et al*., 2023).

Debido a la evidencia sobre las predicciones erróneas del modelo ASFT y sus consecuencias, se decidió diseñar una segunda versión en la cual se excluyeron algunas variables relacionadas con los beneficios públicos que recibían las familias, esto con el objetivo de eliminar sesgos socioeconómicos y raciales que estaban incluidos en dichos registros administrativos (Chouldechova *et al*., 2018). Sin embargo, aunque estas modificaciones representaron algunas mejoras, en la actualidad la herramienta sigue siendo objeto de preocupaciones y críticas debido a que no ha logrado eliminar completamente los sesgos, ya que se siguen utilizando datos como el historial de tratamiento de salud mental, registros educativos y las denuncias relacionadas con el consumo de drogas o alcohol, datos que representan en forma desproporcionada a los grupos marginados, por lo que las conclusiones algorítmicas continúan reflejando las antiguas calificaciones sociales de las familias por parte de la comunidad, las autoridades y los tribunales (Blanchard, 2022).

En septiembre de 2020, el mismo condado de Allegheny implementó un modelo de predicción de riesgos similar al AFST, llamado Hello Baby, con el objetivo de evaluar el riesgo de negligencia y abuso infantil de los recién nacidos. Sin embargo, la investigación de Blanchard (2022) demostró que esta herramienta igualmente presenta sesgos algorítmicos hacia las familias de bajos ingresos y de color, ya que se basa en datos que son tan discriminatorios y sesgados como los datos utilizados por el algoritmo AFST, lo cual genera que muchas familias sean erróneamente investigadas e intervenidas inclusive antes de que ocurra el maltrato o antes de que los recién nacidos abandonen el hospital.



En otros países también se utilizan modelos de predicción de riesgos basados en IA para predecir la negligencia y el abuso infantil; por ejemplo, en 2012 se implementó la herramienta Vulnerable Children Predictive Risk Model en Nueva Zelanda, en 2014 se implementó Early Help Profiling System en el Condado en Londres y en 2018 se implementó el modelo Gladsaxe en Dinamarca. Tal y como describe Lappalainen (2021), estas herramientas también presentan sesgos algorítmicos importantes. En Nueva Zelanda la herramienta fue cuestionada y suspendida porque discriminaba a las familias maoríes y generaba una tasa desproporcionada de separaciones de niños; en Londres se detuvo la implementación del modelo debido a preocupaciones sobre su precisión y falta de transparencia; y en Dinamarca también se detuvo su implementación por cuestiones relacionadas con la vigilancia excesiva en los barrios marginales.

### 3.4. Plataforma Tecnológica de Intervención Social

En 2017 el gobierno de la provincia de Salta, Argentina, anunció un convenio con empresa Microsoft y la Fundación CONIN para la implementación de la Plataforma Tecnológica de Intervención Social, que consiste en un modelo de predicción de riesgo basado en un algoritmo de IA que pronostica el riesgo de deserción escolar y de embarazo en adolescentes. A partir de múltiples variables –asistencias, desempeño académico, origen étnico, situación de discapacidad, número de personas en el hogar, ingresos familiares, nivel educativo del entorno familiar, acceso a servicios básicos

como agua potable y saneamiento, historial de violencia de género intrafamiliar, acceso a servicios de salud sexual y reproductiva–, el algoritmo calcula perfiles de riesgo y clasifica a las adolescentes en diferentes niveles de probabilidad de deserción escolar y embarazo temprano (Pedace *et al*., 2023; Hagerty *et al*., 2023). De acuerdo con sus creadores, los resultados arrojados son utilizados para priorizar intervenciones como tutorías especializadas, apoyo emocional, estímulos económicos, servicios sociales, distribución de anticonceptivos, talleres educativos sobre salud sexual y reproductiva, y apoyo psicológico para las jóvenes identificadas (Ortiz Freuler & Iglesias, 2018).

En 2018 se realizó una prueba piloto del algoritmo en la cual se identificaron 418 niñas y niños con más de 70% de riesgo de deserción escolar y 250 niñas con más de 70% de riesgo de embarazo (Ortiz Freuler & Iglesias, 2018). Aunque en este estudio piloto se reveló que las tasas de falsos positivos fueron relativamente bajas, múltiples especialistas, académicos y organizaciones de la sociedad civil manifestaron serias preocupaciones sobre la posible reproducción de sesgos algorítmicos que pudieran ocasionar la existencia de un trato injusto y discriminatorio hacia las personas de bajos ingresos y minorías étnicas. Específicamente, el estudio de la World Web Foundation destacó la falta de transparencia del algoritmo debido a la falta de accesibilidad de las bases de datos empleadas y el proceso de diseño del modelo (Ortiz Freuler & Iglesias, 2018). Por otro lado, el Laboratorio de Inteligencia Artificial Aplicada (LIAA) de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires (UBA) analizó en profundidad el algoritmo en 2018 y publicó un informe en el que reportaron múltiples problemas como la utilización de datos sesgados e inadecuados y la reutilización del conjunto de entrenamiento como datos de evaluación, lo cual se traduce en resultados que sobredimensionaban del riesgo de deserción escolar y de embarazo en adolescentes (Hagerty *et al*., 2023).



En cuanto a los datos utilizados para entrenar y alimentar el algoritmo, Hagerty *et al*. (2023) y Pedace *et al*. (2023) señalan que estos fueron obtenidos a partir de encuestas realizadas por el gobierno provincial y diversas organizaciones de la sociedad civil. Dichas encuestas se llevaron a cabo entre 2016 y 2017 y abarcaron aproximadamente a 300.000 personas residentes en barrios de bajos ingresos, caracterizados por una alta presencia de comunidades indígenas, como los wichi, kolla y guaraní, así como de migrantes provenientes de Bolivia y Paraguay. Si bien el uso de datos socioeconómicos y demográficos es fundamental para la elaboración de modelos predictivos, su aplicación en este contexto generó preocupaciones significativas, ya que la forma en que se procesan e interpretan estos datos puede conducir a un sesgo algorítmico que resulte en una sobreestimación del riesgo de deserción escolar y embarazo adolescente entre los jóvenes de estas comunidades. Este sesgo no solamente refuerza estereotipos negativos, sino que también contribuye a la estigmatización de sectores vulnerables, al establecer una asociación implícita entre la pobreza, el abandono escolar y la maternidad temprana. Como consecuencia, se corre el riesgo de que las políticas públicas y las intervenciones basadas en estos modelos perpetúen y profundicen la desigualdad y la discriminación que sufren estas comunidades en lugar de abordar sus causas estructurales de manera efectiva.

A pesar de que los funcionarios del Ministerio de Primera Infancia defendieron la neutralidad de la base de datos, argumentando que el algoritmo no presentaba sesgos debido a que su propósito era exclusivamente la predicción dentro de poblaciones vulnerables, su implementación generó un fuerte rechazo por parte de diversas organizaciones de la sociedad civil. En particular, la predicción del embarazo adolescente fue motivo de especial controversia, ya que los colectivos feministas expresaron su oposición, señalando que el algoritmo vulneraba los derechos de niñas y mujeres al ignorar las profundas desigualdades estructurales que enfrentan. Además, alertaron sobre el riesgo de que esta herramienta pudiera ser instrumentalizada por grupos contrarios a los derechos sexuales y reproductivos para deslegitimar la necesidad de leyes que garanticen el acceso al aborto. Esta preocupación se intensificó debido a la participación de la Fundación CONIN en el desarrollo del algoritmo, una organización fundada por el pediatra Abel Albino, conocido por su activismo contra el aborto (Hagerty *et al*., 2023). Debido a estas críticas y al intenso debate público que se generó, el uso del algoritmo para predecir el riesgo de embarazo adolescente se suspendió indefinidamente, mientras que el uso del algoritmo para predecir el riesgo de deserción escolar fue suspendido hasta 2024, cuando se comenzó a implementar formalmente.

A nivel mundial, existen diversas iniciativas tecnológicas similares a la Plataforma Tecnológica de Intervención Social. Un ejemplo notable es el sistema de Alerta Temprana de Deserción Escolar (DEWS, por sus siglas en inglés), implementado en 2012 por el Estado de Wisconsin, Estados Unidos, el cual utiliza un algoritmo de IA y aprendizaje automático para predecir el riesgo de deserción escolar en estudiantes de sexto a noveno grado, permitiendo a las instituciones educativas intervenir oportunamente. Sin embargo, en 2021, un análisis realizado por el Departamento de Instrucción Pública de Wisconsin (DPI, por sus siglas en inglés) evidenció un grave sesgo algorítmico en el DEWS, el cual ocasionaba que el sistema sobreestimara significativamente el riesgo de deserción en estudiantes afroamericanos e hispanos, generando un 42% de falsos positivos en estudiantes de piel negra y un 18% en estudiantes hispanos en comparación con sus pares blancos (Feathers, 2023). Esta distorsión se debe a que el algoritmo utiliza la raza como un factor de predicción, lo que lleva a la estigmatización y discriminación de estos grupos estudiantiles históricamente desfavorecidos.



## 4. Discusión

El análisis de los casos revela que los sesgos algorítmicos presentes en los modelos de predicción de riesgos basados en IA utilizados en la administración pública tienen su origen en el proceso de diseño del algoritmo y en el proceso de entrenamiento. En la fase de diseño, los sesgos pueden surgir de manera accidental debido al tratamiento incorrecto de las variables predictivas de riesgo, o de manera deliberada cuando se seleccionan ciertas variables como predictoras de riesgo debido a los perjuicios y estereotipos presentes en los diseñadores del algoritmo; esto reproduce

las desigualdades y los procesos de exclusión preexistentes en la sociedad. El sesgo por razones de género del sistema COMPAS es un ejemplo de la introducción de un sesgo de forma accidental, ya que el algoritmo fue diseñado para ser neutral en términos de género, lo que llevó a la exclusión de esta variable predictiva. Sin embargo, al no considerar las diferencias en las tasas de reincidencia entre hombres y mujeres, el sistema terminó sobrestimando el riesgo de reincidencia en mujeres. El sesgo por razones étnicas del sistema SyRI es un ejemplo de la introducción deliberada de un sesgo, ya que los diseñadores del algoritmo consideraron que el hecho de tener una nacionalidad diferente a la holandesa era un indicador de riesgo de cometer fraude en las prestaciones sociales, lo que provocó una sobrestimación del riesgo para estos grupos.

Por otro lado, en la etapa de entrenamiento del algoritmo, los sesgos pueden surgir debido al uso de datos históricamente sesgados y a problemas de representatividad de estos datos, lo que refuerza patrones discriminatorios preexistentes, afectando la imparcialidad y la precisión de las decisiones y los resultados generados por el algoritmo. El sesgo socioeconómico del sistema AFST es un ejemplo del uso de datos de entrenamiento históricamente sesgados: debido a que el Departamento de Servicios Humanos interviene con mayor frecuencia en familias de bajos ingresos, el algoritmo se entrena con informes que históricamente han asociado la pobreza con la negligencia y el abuso infantil, lo que resulta en una sobreestimación del riesgo de estas familias. El sesgo socioeconómico de la Plataforma Tecnológica de Intervención Social ejemplifica la sobrerrepresentación de ciertos grupos en los datos de entrenamiento: debido a que las familias de bajos ingresos y las poblaciones indígenas están sobrerrepresentadas en las encuestas utilizadas para construir la base de datos, el algoritmo tiende a sobrestimar el riesgo de deserción escolar y embarazo adolescente en estos grupos.



Los sesgos introducidos en el diseño del algoritmo se refuerzan y amplifican con los sesgos introducidos en su entrenamiento; esta retroalimentación perpetúa y agrava los sesgos algorítmicos, distorsionando aún más los resultados obtenidos. Por ejemplo, el sesgo por razones étnicas del sistema SyRI surgió en el proceso de diseño al utilizar las nacionalidades distintas a la holandesa como un predictor de riesgo; sin embargo, dicho sesgo se agravó al entrenar el algoritmo con datos históricos que ya contenían prejuicios estructurales contra minorías, especialmente descendientes de antiguas colonias holandesas y comunidades de origen turco y marroquí. De esta manera, los algoritmos de los modelos para la predicción de riesgos utilizados en la administración pública pueden ser tan sesgados como las personas que los diseñaron y como los datos que se utilizan para entrenarlos.

Los casos analizados en esta investigación ponen en evidencia de manera contundente los impactos negativos de los sesgos algorítmicos en los modelos de predicción de riesgos empleados en la administración pública. Lejos de ser meros errores técnicos, estos sesgos tienen consecuencias profundas y sistémicas que afectan la vida de muchas personas, ya que no solo llevan a la sobrestimación del

riesgo de comunidades históricamente vulnerables, sino que también profundizan las desigualdades preexistentes, consolidando dinámicas de exclusión, discriminación y marginación que afectan el acceso a derechos fundamentales, programas sociales y servicios públicos esenciales. En el caso del sistema COMPAS, los sesgos algorítmicos afectan profundamente el derecho a la libertad condicional o bajo fianza de afroamericanos, hispanos y mujeres. En el caso del SyRI, los sesgos algorítmicos afectaron el acceso a programas sociales de miles de personas provenientes de comunidades migrantes y de bajos ingresos, y las condujo a situaciones severas de endeudamiento, pobreza, estigmatización social y problemas de salud mental. En el caso del AFST, los sesgos algorítmicos han ocasionado investigaciones injustas y medidas de intervención desproporcionadas hacia familias afroamericanas, multirraciales y de bajos ingresos; inclusive algunas familias han llegado a perder la custodia de sus hijas e hijos. En el caso de la Plataforma Tecnológica de Intervención Social, los sesgos algorítmicos contribuyeron a la estigmatización de miles de adolescentes provenientes de familias indígenas, migrantes y de bajos recursos.

Debido a los preocupantes efectos negativos de los sesgos algorítmicos utilizados en la administración pública, surge una cuestión fundamental que debe ser discutida a profundidad: ¿es realmente posible eliminar por completo los sesgos algorítmicos? Esta pregunta no solamente plantea un desafío técnico, sino que también abre un debate ético, político y social sobre la naturaleza misma de la IA y sus impactos en los procesos de toma de decisiones. Una respuesta factible a este cuestionamiento es que los sesgos algorítmicos en la etapa de diseño se pueden eliminar al diversificar los equipos de desarrollo, incorporando expertos de distintos grupos demográficos para reducir la influencia de perspectivas, prejuicios y sesgos de un grupo particular, promoviendo un diseño más inclusivo y equitativo en los algoritmos (Nadeem *et al*., 2022). Sin embargo, incluso con equipos de diseño con un alto grado de diversidad demográfica, los sesgos pueden persistir debido a estructuras organizacionales, normas culturales o metodologías de desarrollo que privilegian ciertos enfoques y perspectivas (Johansen *et al*., 2020). Otra respuesta factible al anterior cuestionamiento es que los sesgos algorítmicos en la etapa de entrenamiento del algoritmo pueden eliminarse al excluir información sensible de los datos de entrenamiento, como raza, género o nivel socioeconómico; sin embargo, esta estrategia no garantiza la eliminación del sesgo, ya que los algoritmos de aprendizaje automático más sofisticados pueden inferir estas características a través de variables indirectas y reconstruir implícitamente la información oculta, repicando así el sesgo original presente en los datos (Fuchs, 2018).



En este sentido, podemos afirmar que los sesgos algorítmicos son inevitables, ya que los sistemas de IA siempre estarán influenciados por valores, suposiciones, perjuicios y estereotipos de las personas u organizaciones que los crean e implementan, así como de los datos utilizados. En la práctica no existe verdaderamente un algoritmo neutral, ya que cada fase del desarrollo implica decisiones subjetivas que no solo determinan cómo el algoritmo procesa la información, sino también qué realidades prioriza y cuáles ignora. En última instancia, el algoritmo no es una verdad objetiva, sino una construcción estadística con un margen de error inherente, donde los

modelos resultantes reflejan perspectivas humanas codificadas en matemáticas y que generan resultados que son sesgados en menor o en mayor medida (Eubanks, 2017).

Dado que los sesgos son inherentes a los algoritmos, autores como Eubanks (2017), Pedace *et al*. (2023) y Aróstegui (2023) destacan la necesidad de capacitar a los profesionales que utilizan estas herramientas, argumentando que su formación no solo debe enfocarse en la identificación de sesgos, sino también en el desarrollo de habilidades para intervenir activamente en los resultados y decisiones arrojadas por los modelos. De este modo, los profesionales no solo podrán detectar y comprender las limitaciones y desviaciones de los algoritmos, sino que también estarán en condiciones de aplicar estrategias para promover un uso más transparente, ético y equitativo de ellos y mitigar sus efectos negativos. Este enfoque, conocido como *human-in-the-loop*, busca equilibrar la automatización con el juicio humano, permitiendo que los expertos evalúen, ajusten y corrijan predicciones sesgadas antes de que influyan en procesos críticos, como la asignación de recursos o la toma de decisiones en políticas públicas. En este sentido, para mitigar los efectos de los sesgos algorítmicos, es fundamental implementar mecanismos de validación humana en las decisiones automatizadas. Sin embargo, tal y como demuestran los casos analizados en esta investigación, en los cuales los algoritmos funcionan como herramientas de apoyo y sus resultados son evaluados y validados por funcionarios públicos antes de tomar decisiones, este enfoque por sí solo no siempre corrige los sesgos algorítmicos.



En primer lugar, se pueden presentar los sesgos cognitivos que se derivan de la falta de experiencia y conocimiento técnico de los funcionarios públicos y que dificultan su capacidad para evaluar críticamente las decisiones algorítmicas. Debido a esta falta de comprensión, se vuelve más difícil identificar sesgos y corregirlos, lo que perpetúa la aplicación de criterios injustos y discriminatorios (Allhutter *et al*., 2020). Esta situación se presentó en el caso del SyRI, ya que los funcionarios públicos no contaban con los conocimientos y las herramientas necesarias para realizar una auditoría algorítmica, limitando su capacidad para evaluar críticamente los resultados arrojados y ocasionando que la identificación de los sesgos ocurriera después de que el algoritmo ya había impactado negativamente a ciertos grupos (Arts & Van den Berg, 2024).

En segundo lugar, pueden darse los sesgos de automatización que indican una tendencia de los individuos a confiar de manera excesiva en los resultados generados por los algoritmos, ocasionando que los funcionarios acepten sin cuestionamientos las recomendaciones y decisiones algorítmicas, incluso cuando estas presentan sesgos evidentes. De esta manera, en lugar de aplicar un juicio propio y considerar información que contradiga la decisión del algoritmo, pueden validar automáticamente sus sugerencias, reforzando así las desigualdades subyacentes; este fenómeno también se conoce como “tecnochovinismo”: la suposición de que las computadoras son superiores a las personas, o que una solución tecnológica es superior a cualquier otra (Blanchard, 2022). Este escenario se evidenció en el caso del AFST, ya que a pesar de que los funcionarios públicos están capacitados para cuestionar los resultados del algoritmo, rara vez lo hacen, pues depositan su confianza en el sistema y aceptan casi

automáticamente sus predicciones, permitiendo que los sesgos algorítmicos persistan en la toma de decisiones (Eubanks, 2017).

En tercer lugar, existen los sesgos de confirmación, que ocurren cuando los funcionarios públicos tienden a aceptar las decisiones algorítmicas que refuerzan sus propias creencias o prejuicios preexistentes; en lugar de desafiar los resultados del sistema automatizado, pueden usarlos como justificación para sostener decisiones sesgadas, lo que no solo impide la corrección de los sesgos algorítmicos, sino que además contribuye a la amplificación de desigualdades estructurales (Sued, 2022). Este fenómeno se ejemplifica con el caso del COMPAS, donde los jueces y operadores del sistema penal ya tomaban decisiones influenciadas por percepciones y prejuicios raciales sobre la reincidencia delictiva; por lo tanto, el sesgo algorítmico no solo reforzó estas creencias, sino que también las legitimó bajo una apariencia de objetividad tecnológica. Como resultado, la mayoría de las decisiones generadas por el algoritmo son aceptadas sin un análisis crítico, contribuyendo así a perpetuar y profundizar la discriminación en el sistema de justicia penal de Estados Unidos (Angwin *et al*., 2016).

Además de la necesidad de que las decisiones algorítmicas sean validadas por intervención humana, se plantea la importancia fundamental de implementar auditorías algorítmicas como un mecanismo para identificar, analizar y mitigar los sesgos. Uno de los principales obstáculos para la implementación efectiva de auditorías algorítmicas es la falta de transparencia en el funcionamiento de la IA, ya que en muchos casos los algoritmos utilizados en la administración pública son desarrollados por empresas privadas cuyos modelos y códigos fuente están protegidos bajo derechos de propiedad intelectual o industrial. A esto se le conoce como opacidad algorítmica e impide que expertos independientes, reguladores y la sociedad civil puedan acceder, evaluar y cuestionar la forma en que estos sistemas toman decisiones que afectan a millones de personas. Esta situación se ejemplifica claramente en el caso del COMPAS, pues la empresa Northpointe, creadora del algoritmo, se ha negado a dar a conocer el código fuente, argumentando que está protegido por leyes de secreto comercial, impidiendo con ello que se conozca su funcionamiento y dificultando las discusiones legales en torno a su implementación.



Debido a las limitaciones de las estrategias para reducir y mitigar los sesgos algorítmicos en los modelos de predicción de riesgos utilizados en la administración pública, y ante la falta de transparencia que dificulta la realización de auditorías algorítmicas, el rol de la sociedad civil y sus organizaciones es crucial. Los casos analizados en esta investigación muestran que la participación de la sociedad civil no solamente permite evidenciar estos sesgos y sus consecuencias, sino también prevenir su reproducción y promover el desarrollo de sistemas más justos y equitativos. En el caso de sistema COMPAS, el sesgo algorítmico hacia personas afroamericanas fue evidenciado por una investigación de ProPublica, una agencia de noticias independiente y sin fines de lucro. Los sesgos hacia personas hispanas y por razones de género fueron evidenciados por investigaciones académicas independientes, mientras que el American Law Institute ha abogado por la aplicación

de normas diferenciadas para hombres y mujeres con el objetivo de neutralizar los efectos del sesgo de género. En el caso del SyRI, los sesgos fueron evidenciados por un relator especial de la ONU, el cual remitió su informe al Tribunal de La Haya, donde se declaró ilegal el uso del algoritmo. En el caso del AFST, los sesgos han sido evidenciados por investigaciones académicas independientes, lo que ha llevado a varios procesos de reformulación del modelo. Por último, los sesgos en la Plataforma Tecnológica de Intervención Social fueron evidenciados por estudios de la World Web Foundation y del Laboratorio de Inteligencia Artificial Aplicada de la UBA, los cuales fueron retomados por periodistas, organizaciones de la sociedad civil y colectivos feministas, resultando en la suspensión indefinida del algoritmo para predecir el embarazo en adolescentes.

## Conclusiones

En esta investigación se analizó una serie de casos representativos a nivel internacional para identificar y analizar las causas de los sesgos algorítmicos en los modelos de predicción de riesgo utilizados en la administración pública, así como sus principales consecuencias e implicaciones éticas, políticas y sociales. Estos sesgos pueden originarse en el proceso de diseño del algoritmo cuando existen errores en el tratamiento de las variables predictivas de riesgo o cuando se seleccionan ciertas variables como predictoras en función de los perjuicios y estereotipos presentes en los diseñadores del algoritmo. Estos sesgos también se pueden originar en la etapa de entrenamiento del algoritmo cuando se utilizan datos que reflejan perjuicios estructurales y sesgos históricos, o cuando existen problemas de representatividad en dichos datos. Los sesgos algorítmicos ocasionan la sobreestimación del riesgo de grupos sociales históricamente vulnerables, reforzando así las desigualdades y los patrones discriminatorios preexistentes en la sociedad, y consolidando dinámicas de exclusión y marginación que afectan el acceso a derechos fundamentales, programas sociales y servicios públicos esenciales. Cuando los sesgos algorítmicos se filtran en los procesos de toma de decisiones de la administración pública, el problema trasciende lo meramente técnico y se convierte en un desafío ético, político y de justicia social.



Los algoritmos de IA siempre estarán influenciados por los valores y prejuicios de sus creadores y los datos utilizados para su entrenamiento, lo que hace imposible su total neutralidad. Cada decisión en su desarrollo implica subjetividad, afectando qué información se prioriza y cómo se la procesa; esto genera construcciones estadísticas que reflejan sesgos humanos en distintos grados. Los sesgos algorítmicos no se pueden evitar, tan sólo mitigar y regular. Los casos analizados revelan que las estrategias implementadas por las distintas dependencias públicas para mitigar los sesgos algorítmicos resultaron insuficientes. Específicamente, la estrategia implementada para que los funcionarios públicos supervisen y validen las decisiones tomadas por los algoritmos, conocida como *human-in-the-loop*, se mostró ineficaz para mitigar los sesgos observados. Esto se debió a la persistencia

de diversas problemáticas relacionadas con la compleja relación que se establece entre los humanos y los sistemas automatizados, tal y como los sesgos cognitivos, los sesgos de automatización y los de confirmación, los cuales limitan la capacidad de intervención humana para corregir las distorsiones en los resultados algorítmicos.

Ante las fallas en las estrategias de mitigación, los casos analizados demuestran que la participación proactiva de la sociedad civil es fundamental para identificar y visibilizar los sesgos algorítmicos presentes en la administración pública, así como comprender sus implicaciones sociales. Además, su involucramiento no solamente contribuye a prevenir la reproducción de estas desigualdades, sino que también impulsa la creación de mecanismos de control y supervisión que fomenten el desarrollo de sistemas más justos, transparentes y equitativos. La sociedad civil se convierte en un actor clave en la construcción de una IA más ética y alineada con los principios de equidad e inclusión.

Para construir estrategias para la regulación de los sesgos algorítmicos en la administración pública, es fundamental que la implementación de estos sistemas vaya acompañada de mecanismos de transparencia, auditoría y rendición de cuentas, así como de una participación de la sociedad civil y expertos en ética digital, para garantizar que la IA en la administración pública sea una herramienta para la equidad y no un vehículo de exclusión. Diversos países han adoptado enfoques innovadores para garantizar estas condiciones. Por ejemplo, en el año 2020, Uruguay puso en marcha la Estrategia de Inteligencia Artificial para el Gobierno Digital con el propósito de impulsar y consolidar el uso responsable de la IA en la administración pública. Esta estrategia subraya la importancia de que los algoritmos implementados en el sector público respeten los derechos humanos, las libertades individuales y la diversidad. Además, adopta un enfoque de transparencia activa, lo que implica que tanto los algoritmos como los datos utilizados para su entrenamiento, junto con las pruebas y validaciones realizadas, deben estar disponibles para el público, permitiendo así auditorías algorítmicas efectivas. También se establece que cada sistema de IA debe contar con una persona claramente identificada, que asumirá la responsabilidad de sus consecuencias. Por último, la estrategia enfatiza que cualquier dilema ético derivado del uso de IA debe ser analizado y resuelto exclusivamente por seres humanos, garantizando así una toma de decisiones alineada con valores éticos y sociales.



**Reconocimiento de uso de IA**

El autor declara que, una vez concluida la redacción inicial del artículo, se utilizó la herramienta ChatGPT para resumir la información recopilada y analizada sobre los casos de estudio, cuyo desarrollo resultó particularmente extenso. El uso de esta herramienta tuvo como único propósito ajustar la extensión del trabajo al límite máximo de palabras establecido por la normativa editorial de la revista, sin que ello implicara cambios sustantivos al contenido, el análisis crítico realizado, las conclusiones y el uso de fuentes, de las que el autor se hace íntegramente responsable.

**Bibliografía**

Allhutter, D., Cech, F., Fischer, F., Grill, G. & Mager, A. (2020). Algorithmic profiling of job seekers in Austria: How austerity politics are made effective. Frontiers in Big Data, 3, 502780. DOI: https://doi.org/10.3389/fdata.2020.00005.

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016). Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. ProPublica, 23 de mayo. Recuperado de: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Aróstegui, L. A. (2023). El uso de RISCANVI en la toma de decisiones penitenciarias. Estudios Penales y Criminológicos, 44(Ext.), 1-43. DOI: https://doi.org/10.15304/epc.44.8884.

Arts, J. & Van den Berg, M. (2024). What the Dutch benefits scandal and policy's focus on 'fraud' can teach us about the endurance of empire. Critical Social Policy, 45(1), 177-187. DOI: https://doi.org/10.1177/02610183241281346.

Avella, M. D. P. R., Sanabria-Moyano, J. E. & Dinas-Hurtado, K. (2022). Uso del algoritmo COMPAS en el proceso penal y los riesgos a los derechos humanos. Revista Brasileira de Direito Processual Penal, 8(1), 275-310. DOI: https://doi.org/10.22197/rbdpp.v8i1.615.



Baker, R. & Hawn, A. (2022). Algorithmic bias in education. International Journal of Artificial Intelligence in Education, 32, 1052-1092. DOI: https://doi.org/10.1007/s40593-021-00285-9.

Berning, P. A. D. (2023). El uso de sistemas basados en inteligencia artificial por las Administraciones públicas: estado actual de la cuestión y algunas propuestas ad futurum para un uso responsable. Revista de Estudios de la Administración Local y Autonómica, (20), 165-185. DOI: https://doi.org/10.24965/reala.11247.

Blanchard, M. (2022). Predictive Analytics in Child Welfare: Five Principles for Regulating Algorithmic Accountability in a New Wave of Predictive Models. University of Baltimore Law Review, 51(3), 421-448. Recuperado de: https://scholarworks.law.ubalt.edu/ublr/vol51/iss3/5.

Blomberg, T., Bales, W., Mann, K., Meldrum. R. & Nedelec Bales. J. (2010). Validation of the COMPAS risk assessment classification instrument. Tallahassee: Florida State University. Recuperado de: http://criminology.fsu.edu/wp-content/uploads/Validation-of-the-COMPASRisk-Assessment-Classification-Instrument.pdf.

CAF (2021). EXPERIENCIA. Datos e Inteligencia Artificial en el sector público. Caracas: CAF. Recuperado de: https://scioteca.caf.com/handle/123456789/1793.

Chouldechova, A., Benavides-Prado, D., Fialko, O. & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. En Conference on Fairness, Accountability and Transparency (134-148). Recuperado de: https://proceedings.mlr.press/v81/chouldechova18a.html.

Eubanks, V (2017). Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor. Nueva York: St. Martin's Press.

Feathers, T. (2023). False Alarm: How Wisconsin Uses Race and Income to Label Students "High Risk". The Markup, 27 de abril. Recuperado de: https://themarkup.org/machine-learning/2023/04/27/false-alarm-how-wisconsin-uses-race-and-income-to-label-students-high-risk.

Fuchs, D. J. (2018). The dangers of human-like bias in machine-learning algorithms. Missouri S&T's Peer to Peer, 2(1), 1-14. Recuperado de: https://scholarsmine.mst.edu/peer2peer/vol2/iss1/1.

Gerchick, M., Jegede, T., Shah, T., Gutierrez, A., Beiers, S., Shemtov, N., Xu, K., Samant, A. & Horowitz, A. (2023). The devil is in the details: Interrogating values embedded in the allegheny family screening tool. En Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (1292-1310). DOI: https://doi.org/10.1145/3593013.3594081.

Hagerty, A., Aranda, F. & Jemio, D. (2023). Predictive puericulture in Argentina: The Plataforma Tecnológica de Intervención Social and the reproduction of Latin eugenics. BJHS Themes, 8, 221-235. DOI: https://doi.org/10.1017/bjt.2023.2.



Hamilton, M. (2019). The sexist algorithm. Behavioral sciences & the law, 37(2), 145-157. DOI: https://doi.org/10.1002/bsl.2406.

Hamilton, M. (2019). The biased algorithm: Evidence of disparate impact on Hispanics. American Criminal Law Review, 56(4), 1553-1577. Recuperado de: https://www.law.georgetown.edu/american-criminal-law-review/wp-content/uploads/sites/15/2019/06/56-4-the-biased-algorithm-evidence-of-disparate-impact-on-hispanics.pdf.

Johansen, J., Pedersen, T. & Johansen, C. (2020). Studying the transfer of biases from programmers to programs. ArXiv preprint arXiv:2005.08231. DOI: https://doi.org/10.48550/arXiv.2005.08231.

Koniakou, V. (2023). From the "rush to ethics" to the "race for governance" in Artificial Intelligence. Information Systems Frontiers, 25(1), 71-102. DOI: https://doi.org/10.1007/s10796-022-10300-6.

Kordzadeh, N. & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388-409. DOI: https://doi.org/10.1080/0960085X.2021.1927212.

Lappalainen, K. F. (2021). Protecting Children from Maltreatment with the Help of Artificial Intelligence: A Promise or a Threat to Children's Rights? En de Vries, K. & Dahlberg, M. (Eds.), De Lege. Law, AI and Digitalisation (431-466). DOI: https://doi.org/10.33063/dl.vi.433.

Lazcoz Moratinos, G., & Castillo Parrilla, J. A. (2020). Valoración algorítmica ante los derechos humanos y el Reglamento General de Protección de Datos: el caso SyRI. Revista chilena de derecho y tecnología, 9(1), 207-225. DOI: http://dx.doi.org/10.5354/0719-2584.2020.56843.

Marinucci, L., Mazzuca, C. & Gangemi, A. (2023). Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender. AI & Society, 38(2), 747-761. DOI: https://doi.org/10.1007/s00146-022-01474-3.

Martín, J. (2022). Inteligencia artificial, sesgos y no discriminación en el ámbito de la inspección tributaria. Crónica Tributaria, (182/2022), 51-89. DOI: https://doi.org/10.47092/CT.22.1.2.

Nadeem, A., Olivera, M. & Babak, A. (2022). Gender bias in AI-based decision-making systems: a systematic literature review. Australasian Journal of Information Systems, 26, 1-34. DOI: https://doi.org/10.3127/ajis.v26i0.3835.



Ortiz Freuler, J. & Iglesias, C. (2018). Algorithms and Artificial Intelligence in Latin America: A Study of Implementation by Governments in Argentina and Uruguay. Washington DC: World Wide Web Foundation. Recuperado de: https://webfoundation.org/docs/2018/09/WF_AI-in-LA_Report_Screen_AW.pdf.

Pasquinelli, M., Cafassi, E., Monti, C., Peckaitis, H. & Zarauza, G. (2022). Cómo una máquina aprende y falla. Una gramática del error para la Inteligencia Artificial. Revista Hipertextos, 10(17), 13-29. DOI: https://doi.org/10.24215/23143924e054.

Pedace, K., Schleider, T. J. & Balmaceda, T. (2023). Inteligencia artificial y sesgos. El caso de la predicción del embarazo adolescente en Salta. Revista Iberoamericana de Ciencia, Tecnología y Sociedad, 18(53), 9-26. DOI: https://doi.org/10.52712/issn.1850-0013-359.

Peeters, R. & Widlak, A. (2023). Administrative exclusion in the infrastructure- level bureaucracy: The case of the Dutch daycare benefit scandal. Public Administration Review, 83(4), 863-877. DOI: https://doi.org/10.1111/puar.13615.

Red en Defensa de los Derechos Digitales (R3D) (2020). Países Bajos suspende uso de algoritmo por discriminar grupos vulnerables. Red en Defensa de los Derechos Digitales, 17 de febrero. Recuperado de: https://r3d.mx/2020/02/17/paises-bajos-suspende-uso-de-algoritmo-por-discriminar-grupos-vulnerables/.

Rittenhouse, K., Putnam-Hornstein, E. & Vaithianathan, R. (2023). Algorithms, Humans and Racial Disparities in Child Protective Systems: Evidence from the Allegheny Family Screening Tool. Documento de trabajo.

Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M. & Gregor, S. (2022). Algorithmic decision-making and system destructiveness: A case of automatic debt recovery. European Journal of Information Systems, 31(3), 313-338. DOI: https://doi.org/10.1080/0960085X.2021.1960905.

Savage, M. (2021). DWP urged to reveal algorithm that 'targets' disabled for benefit fraud. The Guardian, 21 de noviembre. Recuperado de: https://www.theguardian.com/society/2021/nov/21/dwp-urged-to-reveal-algorithm-that-targets-disabled-for-benefit.

Simon, J., Hang Wong, P. & Rieder. G. (2020). Algorithmic bias and the Value Sensitive Design approach. Internet Policy Review, 9(4), 1-16. DOI: https://doi.org/10.14763/2020.4.1534.

Skeem, J., Monahan, J. & Lowenkamp, C. (2016). Gender, risk assessment, and sanctioning: The cost of treating women like men. Law and Human Behavior, 40(5), 580-593. DOI: https://doi.org/10.1037/lhb0000206.

Sued, G. (2022). Culturas algorítmicas: conceptos y métodos para su estudio social. Revista Mexicana de Ciencias Políticas y Sociales, 67(246), 43-73. DOI: https://doi.org/10.22201/fcpys.2448492xe.2022.246.78422.



Valle-Cruz, D., García-Contreras, R. & Gil-Garcia, J. R. (2024). Exploring the negative impacts of artificial intelligence in government: the dark side of intelligent algorithms and cognitive machines. International Review of Administrative Sciences, 90(2), 353-368. DOI: https://doi.org/10.1177/0020852323118705.

Washington, A. L. (2018). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. Colorado Technology Law Journal, 17, 131-160. Recuperado de: https://ctlj.colorado.edu/wp-content/uploads/2021/02/17.1_4-Washington_3.18.19.pdf.

**Protocolo para investigar desde la IA generativa en ciencias sociales**

**Protocolo para investigar a IA generativa em ciências sociais**

***Protocol for Studies from Generative AI in Social Sciences***

**Lucrecia Agustina Sotelo** *



La inteligencia artificial generativa (IAG) está irrumpiendo en la educación superior con el potencial de revolucionar la enseñanza, el aprendizaje y la investigación. Sin embargo, su integración en la investigación en ciencias sociales, especialmente en la etnografía digital, plantea interrogantes sobre sus limitaciones e implicaciones éticas. Este artículo propone un protocolo para estudios exploratorios utilizando la IAG en ciencias sociales. Primero, argumenta el potencial de la IA para analizar grandes volúmenes de datos, identificar patrones y generar hipótesis. Segundo, destaca la "exploración" como herramienta clave en la etnografía digital. El protocolo incluye identificar el modo de conocer, seleccionar herramientas de IA, recopilar y analizar datos, y generar conclusiones. Se reflexiona sobre su implementación como recurso valioso para investigaciones sistemáticas y rigurosas. Finalmente, se concluye que la IAG, junto al protocolo propuesto, puede transformar la investigación social.

**Palabras clave:** inteligencia artificial; protocolo; ciencias sociales; investigación

* Posdoctora en estudios multidisciplinares, Centro de Estudios Avanzados, Universidad Nacional de Córdoba, Argentina. Docente investigadora de la Universidad Nacional de la Patagonia Austral (UNPA). Coordinadora de la Red de Estudios sobre los Inicios a la Vida Universitaria. Correo electrónico: lucreciasotelo@gmail.com. ORCID: https://orcid.org/0000-0003-0602-5199.

A inteligência artificial generativa (IAG) está a invadir a educação superior com o potencial de revolucionar o ensino, a aprendizagem e a investigação. No entanto, a sua integração na investigação em ciências sociais, especialmente na etnografia digital, levanta questões sobre as suas limitações e implicações éticas. Este artigo propõe um protocolo para estudos exploratórios utilizando a IAG em ciências sociais. Primeiro, argumenta o potencial da IA para analisar grandes volumes de dados, identificar padrões e gerar hipóteses. Segundo, destaca a "exploração" como ferramenta chave na etnografia digital. O protocolo inclui: identificar o modo de conhecer, selecionar ferramentas de IA, recolher e analisar dados e gerar conclusões. Reflete-se sobre a sua implementação como um recurso valioso para investigações sistemáticas e rigorosas. Finalmente, conclui-se que a IAG, juntamente com o protocolo proposto, pode transformar a investigação social.

**Palavras-chave:** inteligência artificial; protocolo; ciências sociais; pesquisa

*Generative artificial intelligence (AI) is breaking into higher education with the potential to revolutionize teaching, learning, and research. However, its integration into social science research, especially in digital ethnography, raises questions about its limitations and ethical implications. This article proposes a protocol for exploratory studies using generative AI in social sciences. Firstly, it considers the potential of AI to analyze large volumes of data, identify patterns, and generate hypotheses. Secondly, it highlights "exploration" as a key tool in digital ethnography. The protocol includes identifying the way of knowing, selecting AI tools, collecting and analyzing data, and generating conclusions. Its implementation is reflected upon as a valuable resource for systematic and rigorous research. Finally, the article concludes that generative AI, together with the proposed protocol, can transform social research.*

***Keywords:*** *artificial intelligence; protocol; social sciences; research*

## Introducción

La inteligencia artificial generativa (IAG) se erige como una fuerza disruptiva en la educación superior. Su facultad para sintetizar contenido novedoso a partir de un *corpus* de datos preexistentes (textos, imágenes, código) inaugura un horizonte de posibilidades para la pedagogía, la didáctica y la investigación. Dicha incursión, no obstante, suscita una serie de desafíos complejos e interrogantes críticos que ponen en tensión las estructuras tradicionales de la universidad (Holmes *et al.*, 2023). Su influencia se extiende desde los paradigmas de la autoría académica hasta los modelos de gestión institucional, lo que confirma su rol como catalizador de un cambio profundo y estructural.

Esta tensión se manifiesta con particular agudeza en el ámbito de las ciencias sociales y, específicamente, en la investigación etnográfica digital. La irrupción de la IAG en este campo impone un doble imperativo: por un lado, explorar su potencial para el análisis de grandes volúmenes de datos cualitativos y la generación de hipótesis (Goyanes & Lopezosa, 2024); por otro, cultivar una alfabetización crítica entre el personal investigador que permita discernir las limitaciones metodológicas y las profundas implicaciones éticas de su aplicación (González-Alcaide, 2024). La IAG, por tanto, no es solo una herramienta, sino un objeto de estudio en sí mismo, un agente que reconfigura las prácticas de producción de conocimiento.

Frente a este panorama, este artículo giro sobre el siguiente problema de investigación: ¿cómo se construye conocimiento válido y riguroso en ciencias sociales a través de la IAG? Para abordar esta cuestión, se adopta un posicionamiento teórico-crítico que enmarca el interrogante en dos perspectivas fundamentales. Primero, desde la mediatización de la cultura (Mata, 2023), se analizará cómo la IAG no solo media, sino que estructura la interacción y el conocimiento. Segundo, desde una teoría crítica de la tecnología (Rosa, 2016; Feenberg, 2012), se buscará trascender la visión instrumentalista del simple "uso" para centrar el análisis en los procesos de "apropiación", donde las prácticas de investigación y la tecnología se transforman mutuamente.



Metodológicamente, la indagación se fundamentará en los "modos de conocer" (Sotelo, 2018): fenomenológico, objetivizante y praxeológico. Este enfoque permitirá deconstruir la noción de sentido común sobre la IAG para desarrollar un protocolo de investigación riguroso. El objetivo final es que dicho protocolo habilite una "descripción densa" (Geertz, 1973) de los fenómenos socioculturales mediados por la IA, explicitando sistemáticamente las decisiones, las tensiones y los debates epistemológicos surgidos tanto del diseño de las herramientas como de su implementación práctica.

Para materializar este enfoque, la investigación se articula en torno a tres objetivos centrales e interconectados. En primer lugar, se propone analizar críticamente el potencial de la IAG en la investigación exploratoria, discerniendo

tanto sus fortalezas para la identificación de patrones y la generación de hipótesis como sus debilidades inherentes. En segundo lugar, y como resultado de dicho análisis, se buscará desarrollar un protocolo específico para la utilización de estas tecnologías en estudios etnográficos digitales, que sirva como una guía rigurosa para la selección de herramientas, el tratamiento de los datos y la interpretación de resultados. Subyacente a todo el proceso, el tercer objetivo consiste en sostener una reflexión sistemática sobre las implicaciones éticas y metodológicas de su uso, abordando cuestiones ineludibles como la privacidad, los sesgos algorítmicos y la transparencia del proceso investigativo.

Este estudio parte de la premisa de que la IAG, si bien puede potenciar la investigación social, requiere de un marco orientador para su aplicación. Se postula que la integración responsable y efectiva de esta tecnología solo es viable a través de un protocolo claro que guíe al personal investigador y de una reflexión crítica constante sobre sus limitaciones y sesgos. En última instancia, este artículo busca ofrecer un aporte sustantivo a la innovación metodológica en las ciencias sociales, proveyendo un marco fundamentado para la apropiación crítica, ética y reflexiva de estas tecnologías emergentes en la construcción de conocimiento.

## 1. Pilares conceptuales para abordar la IAG en la investigación en ciencias sociales



La irrupción de la IAG en las ciencias sociales no representa una ruptura aislada, sino la más reciente manifestación de un proceso histórico más profundo: la mediatización de la cultura (Mata, 2023). Como sus predecesoras tecnológicas, la IAG se integra en la vida cotidiana con la promesa de optimizar y agilizar tareas, permeando las prácticas sociales mientras dialoga con los modos de conocer existentes. Esta dinámica reactualiza tensiones teóricas fundamentales. Al igual que la fotografía cuestionó las nociones de “aura” y “originalidad” (Benjamin, 1936), o que Internet reconfiguró las coordenadas espacio-temporales y aceleró los ritmos sociales (Thompson, 1997; Rosa, 2018), la IAG nos obliga a revisitar el debate sobre la tecnología como mera herramienta o como un agente que co-configura las prácticas en una trama relacional (De Certeau, 1997; Burbules & Callister, 2006).

Desde esta perspectiva crítica, el potencial de la IAG se revela en una doble dimensión: como una herramienta de expansión epistémica y, simultáneamente, como un vector de riesgos socioculturales. Por un lado, sus capacidades para el análisis de datos a gran escala permiten identificar patrones y tendencias antes imperceptibles; facilita la generación de nuevas hipótesis al explorar vastos *corpus* de conocimiento; y optimiza la comunicación científica mediante la asistencia en la creación de contenidos.

No obstante, estas mismas capacidades entrañan desafíos críticos. Su entrenamiento con datos existentes puede reproducir y amplificar sesgos sociales, reforzando así narrativas hegemónicas. Su influencia en la creación de contenido masivo amenaza con incidir en la agenda pública, planteando interrogantes sobre

la soberanía informativa. Asimismo, la aparente facilidad de acceso al conocimiento puede impactar en la formación de una ciudadanía crítica, que requiere habilidades para discernir y contextualizar la información algorítmicamente mediada.

En consecuencia, la pregunta fundamental que emerge no es si se debe usar o no la tecnología, sino más bien: ¿cómo construir conocimiento significativo y crítico en la era de la IAG y la hipermediatización de la cultura? Para abordar este interrogante, es imperativo que la comunidad científica trace caminos que articulen el potencial de esta tecnología con una reflexión ética y responsable, asegurando que su uso promueva la democratización de la información, la pluralidad de voces y el pensamiento crítico.

## 2. Marco metodológico: la complejidad y los modos de conocer en la investigación con IAG

El desafío de construir conocimiento desde la IAG exige un andamiaje metodológico que trascienda el pensamiento lineal y abrace la complejidad. Para comprender las prácticas y significados que emergen en el actual contexto de mediatización (Mata, 2023), es imperativo "romper con el pensamiento lineal [...] para dedicarse a la reconstrucción de las redes de las enmarañadas relaciones que se encuentran presentes en cada uno de los actores" (Bourdieu, 2012, p. 122). Este enfoque relacional es fundamental para analizar un fenómeno como la IAG, donde las dimensiones técnicas, sociales y culturales se encuentran profundamente interdefinidas.



Para operacionalizar esta perspectiva, se adopta la distinción de tres modos de conocimiento (Sotelo, 2018), articulados para el estudio de la IAG.

- El *conocimiento fenomenológico* se aboca a la experiencia vivida, observando cómo la tecnología se integra en las prácticas cotidianas y cómo los sujetos perciben sus posibilidades y limitaciones.
- El *conocimiento objetivizante* se enfoca en desentrañar las estructuras subyacentes que condicionan el desarrollo y la aplicación de la IA Gen, tales como los algoritmos, la lógica de programación y los sesgos inherentes a los datos de entrenamiento.
- El *conocimiento praxeológico* articula los dos niveles anteriores para analizar la relación dialéctica entre las estructuras objetivas y las prácticas de los sujetos, comprendiendo cómo se configuran mutuamente en un proceso continuo.

Es crucial entender que estos tres modos no constituyen etapas secuenciales, sino posibilidades analíticas interrelacionadas que permiten transitar desde la aprehensión del fenómeno en el sentido común hasta la construcción de un conocimiento denso y complejo. Atendiendo a este marco, este artículo detalla el desarrollo de un protocolo de investigación que aplica este enfoque. Dicho protocolo, elaborado en el

seno del Grupo Consolidado “Comunicación, Cultura y Aprendizaje” de la Universidad Nacional de la Patagonia Austral (Argentina), se centra en la aplicación sistemática de los modos fenomenológico y objetivante. El objetivo es sentar las bases para una exploración rigurosa que articule el vasto universo de datos que ofrece la IAG con las prácticas de conocimiento propias de las ciencias sociales.

### 2.1. Fundamentos del protocolo

La propuesta se basa en tres pilares fundamentales:

- *Vigilancia epistemológica:* el protocolo enfatiza la importancia de la vigilancia epistemológica en el uso de IAG. Esto implica un monitoreo constante de los datos generados por la IA, así como una revisión crítica de las fuentes y referencias proporcionadas.
- *IAG como herramienta:* la IAG se utiliza como una herramienta de apoyo a la investigación, no como un sustituto del pensamiento crítico y la reflexión. Las herramientas específicas utilizadas en este protocolo incluyen Gemini Advanced y Notebook LM de Google.

En un enfoque mixto, el protocolo combina dos modos de investigación:



- *Modo fenomenológico:* se centra en la experiencia subjetiva del investigador y en la exploración abierta del tema de investigación.
- *Modo objetivante:* busca identificar patrones, relaciones y estructuras objetivas en los datos generados por la IA.

### 2.2. Pasos para la realización del protocolo

El protocolo se estructura en una serie de pasos que guían la interacción con la IAG:

1. *Selección del tema de investigación:* el tema se elige dentro del área de interés del grupo de investigación, lo que facilita la contextualización y la validación de los resultados.
2. *Construcción del campo semántico:* se utiliza la IAG para explorar y expandir el campo semántico del tema de investigación, identificando conceptos clave, términos relacionados y posibles líneas de indagación.
3. *Formulación de preguntas de investigación:* se elaboran preguntas de investigación abiertas y exploratorias que guíen la interacción con la IAG.
4. *Exploración fenomenológica:* se utiliza la IAG para generar respuestas a las preguntas de investigación y se registra la experiencia subjetiva del investigador

en un cuaderno de campo.

5. *Análisis y reflexión:* se analizan las respuestas generadas por la IA, identificando desafíos, tipos de respuestas y referencias proporcionadas.
6. *Redefinición del campo semántico y las preguntas:* se ajustan el campo semántico y las preguntas de investigación en base a los resultados de la exploración fenomenológica.
7. *Exploración objetivante:* se utiliza la IAG para generar respuestas más estructuradas y objetivas, buscando identificar patrones y relaciones.
8. *Análisis dialéctico:* se examina la interacción entre las estructuras objetivas identificadas y las prácticas subjetivas del investigador, explorando cómo se configuran y transforman mutuamente.
9. *Elaboración del protocolo:* se sistematizan los pasos anteriores en un protocolo de investigación que pueda ser replicado y adaptado a otros contextos.

### 2.3. Desafíos y consideraciones éticas

La aplicación de este protocolo implica considerar los siguientes desafíos y aspectos éticos:

- *Sesgo y calidad de los datos:* es fundamental tener en cuenta el posible sesgo en los datos generados por la IA y evaluar críticamente la calidad de las fuentes y referencias proporcionadas.
- *Transparencia y reproducibilidad:* el protocolo debe ser transparente en cuanto a los métodos utilizados y los resultados obtenidos, y debe permitir la reproducibilidad de la investigación.
- *Implicaciones éticas:* el uso de IAG en la investigación social plantea cuestiones éticas, como la privacidad de los datos, la propiedad intelectual y la responsabilidad del investigador.



Este protocolo ofrece un marco metodológico para la construcción de conocimiento en ciencias sociales utilizando la IAG como herramienta. Se espera que contribuya a la reflexión sobre las posibilidades y los desafíos que implica la integración de esta tecnología en la investigación social.

### 2.4. Construcción del protocolo

A continuación, se detalla la construcción del protocolo a través de su aplicación práctica. El proceso se despliega en una secuencia de pasos iterativos que incluyen: la construcción comparativa del campo semántico y las preguntas de investigación entre la indagación humana y la IAG; una fase de exploración fenomenológica para generar y analizar las respuestas iniciales; y finalmente, una etapa de redefinición y síntesis donde la propia IA se utiliza como herramienta

para consolidar el cuestionario final. Este recorrido ilustra de manera concreta el andamiaje metodológico propuesto.

El diseño y la validación del protocolo se concibieron como un proceso iterativo, materializado durante su aplicación efectiva a un caso de estudio el cual se publicó en una revista académica (Sotelo, 2025). Se detalla dicha implementación en el marco de una investigación en curso, con el fin de ilustrar el funcionamiento del andamiaje metodológico propuesto para la construcción de conocimiento en ciencias sociales. El diseño partió de los siguientes pasos:

- *Selección del tema de investigación:* cartografía de las acciones institucionales para ingreso a la universidad argentina durante el confinamiento por COVID-19.
- *Construcción del campo semántico:* esta acción se llevó a cabo en dos instancias: en primer lugar, se listaron conceptos a partir del tema de investigación; y en segundo lugar, se le solicitó a la IAG que realice un campo semántico del tema. Realizado por la investigadora: acciones institucionales: políticas educativas para el nivel superior durante la pandemia; normativas CIN; normativas SPU; condiciones de accesibilidad a la educación superior: desigualdades educativas; ingreso: acceso, inicio. Realizado por la IAG.



Para la construcción del campo semántico se atendió a los siguientes conceptos centrales, los primeros elaborados por la investigadora los siguientes por la IA:

- *Cartografía:* representación gráfica y análisis de la distribución espacial de las acciones institucionales. Implica una visualización de la información, identificando patrones, concentraciones y vacíos.
- *Acciones institucionales:* medidas, programas, estrategias y recursos implementados por las universidades argentinas para garantizar el ingreso de estudiantes durante el confinamiento por COVID-19.
- *Ingreso a la universidad:* proceso de acceso a la educación superior en Argentina, incluyendo los requisitos, modalidades y procedimientos de admisión.
- *Confinamiento por COVID-19:* periodo de aislamiento social y restricciones a la movilidad implementadas en Argentina para contener la pandemia de COVID-19.

Se refirió a los conceptos relacionados:

- *Universidades argentinas:* instituciones de educación superior públicas y privadas en Argentina.
- *Estudiantes:* individuos que buscan ingresar a la universidad.
- *Modalidad virtual:* educación a distancia mediada por tecnologías digitales.

- *Plataformas digitales:* herramientas tecnológicas utilizadas para la enseñanza y el aprendizaje en línea.
- *Políticas de acceso:* normativas y programas que regulan el ingreso a la universidad.
- *Equidad:* garantizar la igualdad de oportunidades en el acceso a la educación superior.
- *Inclusión:* promover la participación de estudiantes de diversos contextos socioeconómicos y culturales.
- *Brecha digital:* desigualdad en el acceso y uso de las tecnologías digitales.
- *Desigualdad social:* diferencias en las oportunidades y condiciones de vida de las personas.

Por último, se atendió a las siguientes subdimensiones:

- *Tecnologías de la información y la comunicación (TIC):* Internet, computadoras, dispositivos móviles, *software* educativo.
- *Modalidades de ingreso:* cursos virtuales, exámenes en línea, evaluación de portafolios, entrevistas virtuales.
- *Acompañamiento estudiantil:* tutorías virtuales, orientación vocacional en línea, apoyo psicosocial.
- *Recursos de apoyo:* becas de conectividad, préstamo de dispositivos, acceso a *software* y materiales digitales.



El análisis comparativo de los campos semánticos elaborados por la investigadora y la IAG revela diferencias significativas en la profundidad y amplitud con la que se aborda el tema de las acciones institucionales para el ingreso a la universidad argentina durante el confinamiento por COVID-19.

**Tabla 1. Comparación de campos semánticos Investigadora - IA Gen**

| Característica | Campo semántico de la investigadora | Campo semántico de la IAG |
|---|---|---|
| Enfoque | Resumido y específico | Exhaustivo y multidimensional |
| Profundidad | Limitado a conceptos básicos | Profundiza en conceptos, relaciones y subdimensiones |
| Alcance | Centrado en políticas y normativas | Abarca aspectos sociales, tecnológicos y educativos |
| Conceptos centrales | Acciones institucionales, accesibilidad, ingreso | Cartografía, acciones, ingreso, confinamiento, virtualidad |

| | | |
|---|---|---|
| Conceptos relacionados | Desigualdades educativas | Universidades, estudiantes, plataformas, equidad, inclusión, brecha digital, desigualdad social |
| Subdimensiones | Ausentes | TIC, modalidades de ingreso, acompañamiento, recursos |
| Relaciones entre conceptos | No establecidas | Explícitas, conectando conceptos como cartografía y desigualdad |
| Ejemplos concretos | Menciona normativas CIN y SPU | No se limita a ejemplos específicos |

Fuente: elaboración propia.

En virtud de ello se concluye estableciendo las siguientes dimensiones:

- *Amplitud y profundidad:* el campo semántico de la IAG es más amplio y profundo, incluyendo una mayor variedad de conceptos y subdimensiones, lo que permite un análisis más completo del tema.
- *Enfoque holístico:* la IAG adopta un enfoque holístico, considerando no solo las políticas institucionales, sino también los aspectos sociales, tecnológicos y educativos del ingreso a la universidad durante el confinamiento.
- *Énfasis en la virtualidad:* el campo semántico de la IAG refleja la importancia de la modalidad virtual y las tecnologías digitales en el contexto del confinamiento.
- *Cartografía como herramienta:* la IAG introduce el concepto de cartografía como una herramienta para visualizar y analizar la distribución espacial de las acciones institucionales, enriqueciendo el análisis.
- *Consideración de la equidad e inclusión:* la IAG incluye conceptos como equidad, inclusión y brecha digital, permitiendo un análisis más crítico de las acciones institucionales.



En resumen, el campo semántico generado por la IAG es más completo, profundo y orientado a la investigación que el realizado por la investigadora. Esto demuestra el potencial de la IA como herramienta para la exploración y construcción de conocimiento en ciencias sociales. Sin embargo, es importante que se complemente y contextualice la información proporcionada por la IA con su propio conocimiento y experiencia.

- *Formulación de preguntas de investigación.* Al igual que en el punto anterior, la investigadora realiza preguntas en torno al tema y luego le solicitó a la IAG que haga lo propio. ¿Cómo se distribuyen geográficamente las acciones institucionales para el ingreso a la universidad en Argentina durante el confinamiento? ¿Qué tipos de acciones implementaron las universidades

argentinas para garantizar el ingreso en la modalidad virtual? ¿Cómo impactaron las acciones institucionales en el acceso y la permanencia de estudiantes de diferentes contextos socioeconómicos? ¿Qué desafíos enfrentan las universidades argentinas para garantizar la equidad y la inclusión en el ingreso durante el confinamiento?

- *Preguntas de la IAG:* i) Enfoque general: ¿cómo se adaptaron las universidades argentinas al contexto de confinamiento por COVID-19 para garantizar el ingreso de estudiantes? ¿Qué tipo de acciones institucionales se implementaron en Argentina para facilitar el ingreso a la universidad durante la pandemia? ¿Cuáles fueron las principales dificultades que enfrentan las universidades argentinas para garantizar el acceso a la educación superior durante el confinamiento? ii) Enfoque espacial: ¿existe una distribución geográfica desigual de las acciones institucionales implementadas para el ingreso a la universidad durante el confinamiento? ¿Qué factores influyeron en la concentración o dispersión de las acciones institucionales en diferentes regiones de Argentina? ¿Cómo se relaciona la disponibilidad de recursos tecnológicos y de conectividad con la implementación de acciones institucionales en diferentes zonas del país? iii) Enfoque social: ¿cómo impactaron las acciones institucionales en el acceso y la permanencia de estudiantes de diferentes contextos socioeconómicos durante el confinamiento? ¿Se evidenciaron brechas digitales o desigualdades sociales en el acceso a la educación superior durante la pandemia? ¿Qué medidas se implementaron para mitigar las desigualdades y garantizar la inclusión de estudiantes vulnerables durante el confinamiento? iv) Enfoque tecnológico: ¿qué rol desempeñaron las TIC en la implementación de acciones institucionales para el ingreso a la universidad durante el confinamiento? ¿Qué plataformas digitales y herramientas tecnológicas se utilizaron para facilitar el ingreso y el acompañamiento de los estudiantes? ¿Cómo se evaluó la efectividad de las acciones institucionales implementadas en la modalidad virtual? v) Enfoque comparativo: ¿qué diferencias y similitudes se observan en las acciones institucionales implementadas por universidades públicas y privadas durante el confinamiento? ¿Cómo se comparan las experiencias de Argentina con las de otros países de la región en relación con el ingreso a la universidad durante la pandemia? ¿Qué lecciones se pueden aprender de las acciones institucionales implementadas durante el confinamiento para mejorar el acceso a la educación superior en el futuro?



Al comparar las preguntas formuladas por la investigadora y la IAG, se observa que ambas se centran en el tema principal. Sin embargo, la IAG presenta un abanico más amplio y diverso de preguntas, abordando el tema desde diferentes enfoques.

**Tabla 2. Comparación de preguntas de investigación**

| Característica | Preguntas de la investigadora | Preguntas de la IA Gen |
|---|---|---|
| Cantidad | Cuatro preguntas | 15 preguntas |
| Enfoque | Principalmente espacial y social | Abarca enfoques general, espacial, social, tecnológico y comparativo |
| Profundidad | Preguntas generales sobre distribución, tipos de acciones, impacto y desafíos | Preguntas más específicas sobre factores de influencia, brechas digitales, rol de las TIC, efectividad de las acciones, comparaciones entre universidades y países, lecciones aprendidas |
| Ejemplos | "¿Cómo se distribuyen geográficamente las acciones...?" | "¿Qué factores influyeron en la concentración o dispersión de las acciones...?" |

Fuente: elaboración propia.



### 2.5. Análisis

Del análisis de la información se desprenden los siguientes elementos:

- *Mayor amplitud:* la IAG formula una mayor cantidad de preguntas, lo que permite un análisis más exhaustivo del tema.
- *Diversidad de enfoques*: la IAG aborda el tema desde diferentes perspectivas, incluyendo el enfoque espacial, social, tecnológico y comparativo, lo que enriquece la investigación.
- *Mayor profundidad:* las preguntas de la IAG son más específicas y profundizan en aspectos como los factores que influyen en la distribución de las acciones, las brechas digitales, el rol de las TIC, la efectividad de las acciones y las lecciones aprendidas.
- *Potencial para la investigación:* las preguntas de la IAG generan un mayor número de líneas de investigación y permiten abordar el tema con mayor profundidad.

En resumen, las preguntas formuladas por la IAG son más numerosas, diversas y profundas que las planteadas por la investigadora. Esto demuestra la capacidad de la IA para generar preguntas relevantes y estimular la reflexión sobre un tema de investigación.

### 2.6. Exploración fenomenológica

Para dar cuenta de este paso, se tomaron todas las preguntas, tanto las realizadas por la investigadora como por la IAG. Estas últimas se las colocaron como intenciones al Gemini Advanced con el siguiente enunciado: "Responde esta pregunta", dos puntos y la pregunta a continuación. A los efectos de este artículo solo se presentan el abordaje de dos preguntas, una para las realizada por la investigadora y otra por la IAG. Sin embargo, las dos responden a la misma dimensión.

A la pregunta de la investigadora "¿Cómo se distribuyen geográficamente las acciones institucionales para el ingreso a la universidad en Argentina durante el confinamiento?", se obtuvo la siguiente respuesta:

> "Para responder a la pregunta sobre la distribución geográfica de las acciones institucionales para el ingreso a la universidad en Argentina durante el confinamiento por COVID-19, es necesario analizar diversos factores y considerar la heterogeneidad del sistema universitario argentino.
> • Aspectos a considerar:
> • Contexto socioeconómico**:** Las universidades ubicadas en regiones con mayores recursos económicos y mejor infraestructura tecnológica pudieron implementar acciones más completas y llegar a un mayor número de estudiantes.
> • Conectividad: La brecha digital entre zonas urbanas y rurales influyó en la accesibilidad a las acciones implementadas en la modalidad virtual.
> • Tipo de universidad: Las universidades públicas, con mayor matrícula y diversidad de estudiantes, enfrentaron mayores desafíos para garantizar la equidad y la inclusión en el acceso.
> • Políticas institucionales: Cada universidad implementó diferentes estrategias y acciones, dependiendo de sus recursos, prioridades y características.
> • Posibles patrones de distribución:
> • Concentración en zonas urbanas**:** Es probable que las acciones se hayan concentrado en las principales ciudades y zonas con mayor conectividad.
> • Dispersión en universidades públicas**:** Las universidades públicas, con presencia en diversas regiones del país, pudieron implementar acciones en un territorio más amplio.
> • Heterogeneidad en la oferta: La diversidad de acciones implementadas pudo generar diferencias en el acceso a la educación superior en diferentes zonas.
> • Para una respuesta más precisa, se requiere:
> • *Cartografía de las acciones:* Elaborar un mapa que muestre la ubicación de las universidades y las acciones implementadas en cada una.

> • *Análisis de datos:* Recopilar y analizar datos sobre la cantidad de estudiantes que accedieron a las acciones en diferentes regiones.
> • *Estudios de caso:* Investigar en profundidad las experiencias de diferentes universidades y estudiantes en distintos contextos.
> • Ejemplos de acciones:
> • *Plataformas virtuales*: Muchas universidades implementaron plataformas virtuales para ofrecer cursos de ingreso, tutorías y acompañamiento a los estudiantes.
> • *Becas de conectividad*: Se otorgaron becas para que los estudiantes puedan acceder a internet y participar en las actividades virtuales.
> • *Exámenes en línea:* Se adaptaron los exámenes de ingreso a la modalidad virtual.
> • *Tutorías virtuales:* Se brindaron tutorías en línea para apoyar a los estudiantes en el proceso de ingreso.
> En conclusión, la distribución geográfica de las acciones institucionales para el ingreso a la universidad en Argentina durante el confinamiento es un tema complejo que requiere un análisis detallado de diversos factores. Es probable que se observe una concentración de acciones en zonas urbanas y una mayor dispersión en universidades públicas, pero también es esperable encontrar una gran heterogeneidad en la oferta de acciones en diferentes regiones del país".



A la pregunta de la IAG "¿Cómo se adaptaron las universidades argentinas al contexto de confinamiento por COVID-19 para garantizar el ingreso de estudiantes?", se obtuvo la siguiente respuesta:

> "Las universidades argentinas demostraron una gran capacidad de adaptación durante el confinamiento por COVID-19 para garantizar el ingreso de estudiantes. Se implementaron diversas estrategias y acciones que transformaron el proceso de admisión, impulsando la virtualidad y la innovación educativa.
> Principales adaptaciones:
> • Virtualización del proceso de ingreso:
> - Cursos de ingreso virtuales: Se migraron los cursos de ingreso a plataformas digitales, utilizando herramientas como videoconferencias, foros de discusión y materiales interactivos.
> - Exámenes en línea: Se implementaron exámenes de ingreso en línea, con diferentes modalidades de evaluación y sistemas de *proctoring* para garantizar la integridad académica.
> - Inscripciones online: Se facilitó la inscripción a través de plataformas digitales, evitando trámites presenciales y agilizando el proceso.
> • Fortalecimiento de la infraestructura tecnológica:

- Plataformas de aprendizaje: Se invirtió en plataformas de aprendizaje más robustas y con mayor capacidad para alojar a un gran número de estudiantes.
- Capacitación docente: Se brindó capacitación a los docentes en el uso de herramientas digitales y en metodologías de enseñanza virtual.
- Apoyo técnico: Se fortaleció el soporte técnico para asistir a los estudiantes y docentes con las dificultades técnicas.

• Acciones para mitigar la brecha digital:
  • Becas de conectividad: Se otorgaron becas para que los estudiantes puedan acceder a internet y participar en las actividades virtuales.
  • Préstamo de dispositivos: Se implementaron programas de préstamo de computadoras y tablets a estudiantes que no contaban con los recursos necesarios.
  • Espacios de acceso a internet: Se habilitaron espacios con conectividad en las universidades para que los estudiantes puedan acceder a internet y realizar sus actividades.
• Acompañamiento estudiantil:
  - Tutorías virtuales: Se brindaron tutorías en línea para apoyar a los estudiantes en el proceso de ingreso y durante el primer año de cursada.
  - Orientación vocacional online: Se ofrecieron servicios de orientación vocacional en línea para ayudar a los estudiantes a elegir su carrera.
  • Apoyo psicosocial: Se brindó apoyo psicosocial a los estudiantes para ayudarlos a afrontar las dificultades emocionales del confinamiento y la adaptación a la virtualidad.
• Flexibilización de requisitos:
  • Prórrogas para la presentación de documentación: Se extendieron los plazos para la presentación de documentación requerida para el ingreso.
  • Modalidades de evaluación alternativas: Se implementaron modalidades de evaluación más flexibles, como la evaluación de portafolios o entrevistas virtuales.

Desafíos:

• Garantizar la equidad y la inclusión: Asegurar que todos los estudiantes, independientemente de su contexto socioeconómico y su ubicación geográfica, tengan las mismas oportunidades de acceder a la educación superior.
• Mantener la calidad educativa: Asegurar que la modalidad virtual no afecte la calidad de la enseñanza y el aprendizaje.
• Evaluar la efectividad de las acciones: Medir el impacto de las acciones implementadas y realizar los ajustes necesarios.

En general, las universidades argentinas lograron adaptarse al contexto de confinamiento y garantizar el ingreso de estudiantes,

> impulsando la virtualidad y la innovación educativa. Sin embargo, es necesario seguir trabajando para superar los desafíos pendientes y garantizar una educación superior inclusiva y de calidad para todos".

Del cotejo entre ambas respuestas emergen dimensiones de análisis que revelan dos lógicas de producción de conocimiento distintas, aunque complementarias. El abordaje derivado del protocolo de investigación se caracteriza por la cautela epistemológica y una orientación hacia la profundidad analítica. Desde esta perspectiva, se subraya la complejidad del fenómeno y la necesidad de futuras indagaciones –como cartografías o estudios de caso– para formular conclusiones robustas, utilizando un lenguaje deliberadamente probabilístico.

En contraste, la IAG opera desde una lógica de amplitud descriptiva y concreción fáctica. Su respuesta, de tono más optimista, prioriza la enumeración de un vasto repertorio de acciones ya implementadas por las instituciones, ofreciendo un panorama amplio sin detenerse en el análisis de los factores o patrones subyacentes.

En síntesis, se contrapone un enfoque analítico y reflexivo, propio de la indagación científica, con uno descriptivo y propositivo, característico del procesamiento algorítmico de la información. Lejos de ser excluyentes, ambas perspectivas aportan dimensiones valiosas para la comprensión del objeto de estudio.



**Tabla 3. Análisis comparativo de respuestas**

| Característica | Respuesta de la investigadora | Respuesta de la IA Gen |
|---|---|---|
| Enfoque | Analítico y reflexivo | Descriptivo y propositivo |
| Profundidad | Se centra en la complejidad del tema y la necesidad de más investigación | Detalla las acciones concretas implementadas por las universidades |
| Estructura | Plantea aspectos a considerar, posibles patrones y métodos para obtener una respuesta precisa | Describe las principales adaptaciones, los desafíos y concluye con una valoración general |
| Ejemplos | Menciona ejemplos generales de acciones | Proporciona ejemplos específicos de acciones |
| Tono | Cauteloso y explícito en las limitaciones | Más asertivo y optimista |

Fuente: elaboración propia.

Siendo así, se utiliza la IAG para generar respuestas a las preguntas de investigación y se registra la experiencia subjetiva del investigador en un cuaderno de campo. A partir de ello, se redefinen el campo semántico y las preguntas. De esta manera, los análisis comparativos se reajustaron las preguntas de investigación integrándose en un mismo cuestionario. Para ello, se le pidió al Gemini Advanced que realice la siguiente intención: "Agrupa las preguntas en dimensiones integrando las que son similares". Su respuesta fue la siguiente:

> "1. Adaptación al contexto de confinamiento:
> • ¿Cómo se adaptaron las universidades argentinas al contexto de confinamiento por COVID-19 para garantizar el ingreso de estudiantes?
> • ¿Qué tipo de acciones institucionales se implementaron en Argentina para facilitar el ingreso a la universidad durante la pandemia?
> • ¿Cuáles fueron las principales dificultades que enfrentan las universidades argentinas para garantizar el acceso a la educación superior durante el confinamiento?
> • 2. Distribución geográfica y equidad:
> • ¿Cómo se distribuyen geográficamente las acciones institucionales para el ingreso a la universidad en Argentina durante el confinamiento?
> • ¿Existe una distribución geográfica desigual de las acciones institucionales implementadas para el ingreso a la universidad durante el confinamiento?
> • ¿Qué factores influyeron en la concentración o dispersión de las acciones institucionales en diferentes regiones de Argentina?
> • ¿Cómo se relaciona la disponibilidad de recursos tecnológicos y de conectividad con la implementación de acciones institucionales en diferentes zonas del país?
> • ¿Qué desafíos enfrentan las universidades argentinas para garantizar la equidad y la inclusión en el ingreso durante el confinamiento?
> 3. Impacto en el acceso y la permanencia:
> • ¿Cómo impactaron las acciones institucionales en el acceso y la permanencia de estudiantes de diferentes contextos socioeconómicos?
> • ¿Cómo impactaron las acciones institucionales en el acceso y la permanencia de estudiantes de diferentes contextos socioeconómicos durante el confinamiento?
> • ¿Se evidenciaron brechas digitales o desigualdades sociales en el acceso a la educación superior durante la pandemia?
> • ¿Qué medidas se implementaron para mitigar las desigualdades y garantizar la inclusión de estudiantes vulnerables durante el confinamiento?

4. Rol de las TIC y la virtualidad:
• ¿Qué tipos de acciones implementaron las universidades argentinas para garantizar el ingreso en la modalidad virtual?
• ¿Qué rol desempeñaron las TIC en la implementación de acciones institucionales para el ingreso a la universidad durante el confinamiento?
• ¿Qué plataformas digitales y herramientas tecnológicas se utilizaron para facilitar el ingreso y el acompañamiento de los estudiantes?
• ¿Cómo se evaluó la efectividad de las acciones institucionales implementadas en la modalidad virtual?
5. Comparación y aprendizajes:
• ¿Qué diferencias y similitudes se observan en las acciones institucionales implementadas por universidades públicas y privadas durante el confinamiento?
• ¿Cómo se comparan las experiencias de Argentina con las de otros países de la región en relación con el ingreso a la universidad durante la pandemia?
• ¿Qué lecciones se pueden aprender de las acciones institucionales implementadas durante el confinamiento para mejorar el acceso a la educación superior en el futuro?".



Esta unificación de preguntas permite una visión más organizada y completa de las dimensiones a investigar sobre el tema de las acciones institucionales para el ingreso a la universidad en Argentina durante el confinamiento por COVID-19. Para continuar con el análisis, y con ello unificar respuestas, resulta de vital importancia construir las referencias textuales para así cotejar si la información. En este sentido, se le solicitó que construya las referencias y que las construya en formato APA 7. Para ejemplificar esto se presentan las referencias de la respuesta anterior.

"Referencias:
• Consejo Interuniversitario Nacional (CIN). (2020). *Las universidades argentinas frente a la pandemia del COVID-19*. Enlace al documento
• Bnm.me.gov.ar. (2021). *Nuestras universidades públicas argentinas frente a la pandemia COVID-19*. Enlace al documento
• Este libro analiza las estrategias implementadas por las universidades públicas argentinas para afrontar los desafíos de la pandemia, incluyendo la virtualización de la enseñanza y el apoyo a los estudiantes.
• Ri.conicet.gov.ar. (2021). *Nuestras universidades públicas argentinas frente a la pandemia COVID-19*. Enlace al documento
• Esta publicación del CONICET examina las respuestas de

las universidades públicas a la pandemia, con un enfoque en las desigualdades y la inclusión.
• Sobre la virtualización de la educación superior:
• García Aretio, L. (2014). *Educación a distancia hoy*. Madrid: Síntesis.
• Este libro ofrece una perspectiva general sobre la educación a distancia, incluyendo las modalidades, las tecnologías y los desafíos.
• Salinas, J. (2004). *Innovación docente y uso de las TIC en la enseñanza universitaria*. Revista Universidad y Sociedad del Conocimiento, 1(1). 1. revistas.uminho.pt revistas.uminho.pt
• Este artículo analiza la innovación docente y el uso de las TIC en la enseñanza universitaria, temas relevantes en el contexto de la virtualización del ingreso.
• Sobre la equidad y la inclusión en la educación superior:
• Dussel, I. (2010). *Aprender y enseñar en la cultura digital*. Buenos Aires: Santillana.
• Este libro aborda las implicancias de la cultura digital en la educación, incluyendo las brechas digitales y la necesidad de garantizar la inclusión.
• Tedesco, J. C. (2005). *Educar en la sociedad del conocimiento*. Buenos Aires: Fondo de Cultura Económica.
• Este libro analiza los desafíos de la educación en la sociedad del conocimiento, con un enfoque en la equidad y la inclusión".



### 2.7. Exploración objetivante

La IAG puede construir informes dado que responde a la pregunta que se le realiza. Ello conduce a dos instancias importantes: i) cómo se construye la pregunta; y ii) se debe considerar que, al ser un "tipo de lenguaje", puede nutrirse de textos –en todos los formatos posibles– que muchas veces no tienen referencias académicas. Para generar respuestas más estructuradas y objetivas, es importante conocer el tema que se investiga en profundidad. Siendo así, sin marco teórico la identificación de patrones y relaciones, puede quedar a un nivel muy superficial.

En un análisis dialéctico, se examinaron las interacciones entre las estructuras objetivas identificadas y las prácticas subjetivas de la investigadora. Se llegó a la conclusión que la relación con la IAG favorece la construcción dialógica, dado que, al seguir el esquema planteado, las preguntas y las respuestas se enriquecen.

## 3. Elaboración del protocolo

A partir del análisis llevado a cabo se diseña el siguiente protocolo para la construcción de conocimiento en ciencias sociales:

a. *Objetivo:* sistematizar la interacción con la IAG para la construcción de conocimiento en ciencias sociales.
b. *Herramientas:*
   • IAG (por ejemplo: Gemini Advanced)
   • Cuaderno de campo
c. *Pasos:*
   1. Selección del tema de investigación:
   - Elegir un tema dentro del área de interés del investigador.
   - Formular una pregunta de investigación inicial.
   2. Construcción del campo semántico:
   - Fase 1. Elaboración propia: el investigador realiza un listado de conceptos clave relacionados con el tema.
   - Fase 2. Generación con IA: solicitar a la IAG que construya un campo semántico del tema.
   - Fase 3. Comparación y análisis: comparar ambos campos semánticos, identificando diferencias en profundidad, amplitud y enfoque. Analizar las fortalezas y debilidades de cada uno.
   3. Formulación de preguntas de investigación:
   - Fase 1. Elaboración propia: el investigador formula preguntas de investigación en torno al tema.
   - Fase 2. Generación con IA: solicitar a la IAG que formule preguntas de investigación sobre el tema.
   - Fase 3. Comparación y análisis: comparar las preguntas formuladas, analizando la cantidad, el enfoque, la profundidad y el potencial para la investigación.
   4. Exploración fenomenológica:
   - Ingresar las preguntas de investigación en la IAG.
   - Registrar las respuestas generadas por la IA en un cuaderno de campo.
   - Reflexionar sobre la experiencia de interacción con la IA, las respuestas obtenidas y las ideas que surgen.
   5. Redefinición del campo semántico y las preguntas:
   - Ajustar el campo semántico y las preguntas de investigación en base a los resultados de la exploración fenomenológica.
   - Integrar las preguntas en un mismo cuestionario, agrupándolas por dimensiones.
   6. Exploración objetivante:
   - Formular preguntas más específicas y orientadas a obtener respuestas estructuradas y objetivas.
   - Considerar el marco teórico y el estado del arte sobre el tema para formular preguntas relevantes.
   - Analizar las respuestas generadas por la IA, identificando patrones, relaciones y estructuras objetivas.
   7. Análisis dialéctico:
   - Examinar las interacciones entre las estructuras objetivas identificadas y las prácticas subjetivas del investigador.
   - Analizar cómo se configuran y transforman mutuamente las perspectivas del investigador y la IA.

- Reflexionar sobre el proceso de construcción de conocimiento en diálogo con la IA.

8. Referencias bibliográficas:

- Solicitar a la IAG que genere referencias bibliográficas relevantes para el tema en formato APA 7.
- Revisar y completar las referencias con otras fuentes académicas.

d. *Consideraciones:*

• Vigilancia epistemológica: mantener una actitud crítica durante todo el proceso, evaluando la calidad y el sesgo de la información generada por la IA.
• Pensamiento crítico: utilizar la IA como herramienta de apoyo, sin sustituir el pensamiento crítico y la reflexión propia.
• Transparencia: documentar el proceso de investigación, incluyendo las interacciones con la IA y las decisiones tomadas.

Este protocolo busca guiar la investigación con IAG en ciencias sociales, promoviendo un enfoque crítico y reflexivo en la construcción de conocimiento.

## 4. **Resultados**, **análisis y discusión**

El acercamiento a la IAG en este artículo no partió de una necesidad instrumental, sino de una exploración metodológica para evaluar su rol en la construcción de conocimiento. Dicha exploración arrojó una serie de reflexiones críticas sobre su potencial y sus limitaciones en el campo de las ciencias sociales.



En primer lugar, se constató que la aparente “intuitividad” de herramientas como Gemini Advanced, si bien facilita su apropiación técnica, exige como contraparte una planificación investigativa aún más rigurosa. La eficacia de la IA no reside en la herramienta en sí, sino en la solidez del marco teórico y la claridad de los objetivos que orientan su aplicación. Sin una definición precisa del problema, la exploración deviene en un ejercicio estéril.

En segundo lugar, se confirmó el considerable potencial de la IAG para el análisis de grandes volúmenes de datos, permitiendo abordar la complejidad del contexto de una manera más eficiente. Un ejemplo de ello fue su utilidad para reconstruir el *corpus* de políticas implementadas por las universidades argentinas durante la pandemia, proveyendo un marco contextual robusto para el análisis. No obstante, este potencial para el procesamiento de datos debe ser matizado en su aplicación a la escritura académica. Si bien la IA puede asistir en la forma, el proceso de construcción de sentido permanece como una labor intransferible del investigador. Su uso exige, por tanto, una posición crítica que evite la dependencia y promueva la autonomía intelectual.

Finalmente, la experiencia reafirmó que el deseo y la curiosidad continúan siendo el motor irreductible de la investigación, incluso en un entorno de creciente

automatización. La IAG se perfila como una herramienta prometedora para la etnografía digital, habilitando abordajes macro que antes eran inviables. Sin embargo, esta capacidad para la amplitud no suplanta, sino que debe ser articulada con los pilares de la investigación social: la interpretación densa de las culturas (Geertz, 1973) y el análisis de la complejidad (García, 2006) siguen siendo fundamentales en la era de la IA.

## Conclusión: hacia una investigación social aumentada por la IAG

A lo largo de este artículo se ha buscado responder al interrogante sobre cómo construir conocimiento válido y riguroso en ciencias sociales a través de la IAG. La respuesta que emerge de esta exploración metodológica es que dicha construcción es posible, no mediante la sustitución de la labor investigativa, sino a través de una alianza cognitiva estructurada y críticamente consciente entre la capacidad computacional de la IA y la sensibilidad interpretativa del investigador.

En cumplimiento del primer objetivo, se analizó el potencial de la IAG, constatando su notable capacidad para expandir el campo semántico y formular un abanico de preguntas de investigación con mayor amplitud y diversidad que el enfoque humano inicial. No obstante, se evidenció que su producción tiende a una lógica descriptiva y propositiva, carente del análisis y la cautela epistemológica que caracterizan a la indagación científica. Para dar respuesta al segundo objetivo, se desarrolló y validó un protocolo de investigación que sistematiza la interacción con la IA. Este protocolo, que articula los modos de conocer fenomenológico y objetivisante en un proceso iterativo, se erige como la principal contribución de este artículo. Su aplicación demuestra que es posible encauzar el potencial de la IA, transformando la exploración inicial en un diálogo estructurado que profundiza la investigación. Finalmente, atendiendo al tercer objetivo, la implementación del protocolo ha permitido reflexionar sistemáticamente sobre las implicaciones éticas y metodológicas de este nuevo escenario. Se concluye que la vigilancia epistemológica, la conciencia sobre los sesgos algorítmicos y la transparencia del proceso no son consideraciones adyacentes, sino condiciones de posibilidad para una investigación social aumentada por la IA que sea rigurosa y pertinente.



En definitiva, este artículo sostiene que el futuro de la investigación social no reside en una disyuntiva entre la metodología tradicional y la inteligencia artificial, sino en su integración dialéctica. El protocolo aquí presentado ofrece un camino para navegar esta nueva frontera, promoviendo una apropiación crítica y reflexiva de la tecnología que, lejos de amenazar la integridad de nuestra disciplina, potencia su capacidad para comprender la complejidad del mundo contemporáneo.

## Aclaración

Este artículo forma parte de una reflexión metodológica dentro de un proyecto de investigación 29/D117: “Cartografías sobre los ingresos a la universidad pública Argentina en tiempo de confinamiento por COVID-19”.

## Reconocimiento de uso de IA

A modo de nota metodológica, se informa que para la elaboración del artículo se ha empleado IAG como tecnología de asistencia. La autora, quien presenta una condición de neurodivergencia (dislexia), utilizó el modelo Gemini Advanced bajo un protocolo de trabajo iterativo: tras una redacción inicial propia, se aplicó la herramienta para la detección y corrección de patrones de error específicos (“marcas disléxicas”), seguida de una validación y ajuste final por parte de la autora para garantizar la fidelidad al pensamiento original. Este procedimiento, destinado a mejorar la claridad del texto atendiendo a una discapacidad, subraya el potencial de la IA como herramienta para la inclusión en la comunicación científica.

## Bibliografía



Amozurrutia, J. A. (2011). Complejidad y Ciencias Sociales. Un modelo adaptativo para la investigación interdisciplinaria. México: UNAM.

Feenberg, A. (2012). Transformar la tecnología. Una nueva visita a la teoría crítica. Bernal: Universidad de Quilmes.

Garcia Peñalvo, F. J. (2024). Inteligencia Artificial Generativa: un análisis desde múltiples perspectivas. Education in the Knowledge Society, 25. DOI: https://doi.org/10.14201/eks.31942.

González Alcaide, G. (s/f). Inteligencia artificial generativa: Un contexto disruptivo en el acceso a la información. Infonomy, 2(1), e24013. DOI: https://doi.org/10.3145/infonomy.24.013.

Goyanes, M. & Lopezosa, C. (2024). ChatGPT en educación: Usos, retos y oportunidades de la Inteligencia Artificial Generativa en Educación. Buenos Aires: GEDISA.

Guersenzvaig, A., Sánchez-Monedero, J., Picassó i Piquer, M. & Gopegui, B. (s/f). Inteligencia artificial generativa en la educación universitaria: la urgencia de una perspectiva crítica. Daimon. Revista Internacional de Filosofía, en prensa. DOI: http://dx.doi.org/10.6018/daimon.636721.

Mata, M. M. (2023). Indisciplinada. Buenos Aires: Fundación Friedrich Ebert.

Pichardo, J. I., Borrás-Gené, O., Santoro Domingo, P., Martínez Álvaro, L. & Carabantes-Alarcó, D. (2024). Inteligencia artificial generativa como recurso en los procesos de enseñanza y aprendizaje en educación superior. En J. L. Ayala Rodrigo (Coord.), III Jornada «Aprendizaje Eficaz con TIC en la UCM» (33-46). DOI: https://dx.doi.org/10.5209/act.001.

Rosa, H. (2016). Alienación y aceleración. Hacia una teoría crítica de la temporalidad en la modernidad tardía. Buenos Aires: Katz.

Rosa, H. (2020). Resonancia. Una sociología de la relación con el mundo. Buenos Aires: Katz.

Sotelo, L. (2018). Estudiar en los confines. Río Gallegos: UNPAEdita.

Sotelo, L (2025) Desigualdad digital y virtualización forzada: el ingreso a la universidad pública argentina en pandemia. Revista IRICE, (48), e2030. Recuperado de: https://ojs.rosario-conicet.gov.ar/index.php/revistairice/article/view/2030.

# Comité de lectura



Los siguientes expertos integraron el comité de lectura que se ocupó de considerar, evaluar y seleccionar los artículos académicos en el marco de la convocatoria abierta a artículos sobre inteligencia artificial y ética que llevaron adelante la Organización de Estados Iberoamericanos (OEI) y la Pontificia Comisión para América Latina (PCAL):

**Ismael Gómez**: con un origen universitario en filosofía, posee un máster en tecnología de la información y la comunicación aplicadas a la educación y formación por la Universidad Autónoma de Madrid y otro por la Universidad de Salamanca, ambas de España. Emprendedor en el mundo editorial y autor de diversos artículos relacionados con la tecnología y la digitalización. Desde 2023 es el director de estrategia digital global de la Organización de Estados Iberoamericanos (OEI).

**Diego Álvarez Newman**: sociólogo, profesor en enseñanza media y superior, y doctor en ciencias sociales de la Universidad de Buenos Aires (UBA), Argentina. Se desempeña como investigador del CONICET en el Instituto de Estudios Sociales en Contextos de Desigualdades (IESCODE), de la Universidad Nacional de José Clemente Paz (UNPAZ), y como coordinador del Grupo de Trabajo CLACSO "Transiciones justas y cuidado de la casa común". Sus líneas de investigación están centradas en la organización y gestión del trabajo, las subjetividades laborales, las políticas públicas para la inclusión laboral y la plataformización del trabajo.

**Betty Estévez Cedeño**: profesora ayudante doctora en el Departamento de Sociología y Antropología de la Universidad de La Laguna (ULL), España. Obtuvo la licenciatura en sociología en la Universidad Central de Venezuela. Realizó una especialización en gestión de la ciencia y la tecnología en la Universidad Carlos III de Madrid, España, y se doctoró en 2010 en el programa Filosofía, Ciencia, Tecnología y Valores en la Universidad del País Vasco (UPV/EHU), España. Ha sido investigadora predoctoral en el Instituto de Filosofía del Consejo Superior de Investigaciones Científicas (CSIC), y ha trabajado como gestora de proyectos de I+D en la Universidad Complutense de Madrid y en la Universidad Carlos III de Madrid y como técnica de proyectos en la Fundación Española para la Ciencia y Tecnología (FECYT) y el CIEMAT. Ha ejercido como docente en la UNIR y en la Fundación General de la ULL. Entre 2019 y 2020 formó parte del gabinete de la Consejería de Educación del Gobierno de Canarias, contribuyendo a la gestión de las políticas públicas en materia educativa. Sus líneas de investigación se orientan al ámbito educativo, el impacto de la tecnología en la educación, grupos vulnerables y género.

**Luis O. Jiménez Rodríguez**: miembro de la Compañía de Jesús desde 2003. Tiene un doctorado en ingeniería eléctrica de la Universidad de Purdue en Indiana, Estados Unidos. Es ingeniero profesional miembro del Colegio de Ingenieros y Agrimensores de Puerto Rico. Fue profesor del Departamento de Ingeniería Eléctrica y de Computadoras de la Universidad de Puerto Rico en Mayagüez por 23 años. Durante ese tiempo desarrolló, junto a un equipo de profesores y estudiantes, algoritmos de inteligencia artificial discriminativa. En 2012 obtuvo su doctorado en teología fundamental de la Universidad Católica de Lovaina en Bélgica, recibiendo la más alta distinción. Actualmente es catedrático y director de la recién fundada Escuela de Teología de la Pontificia Universidad Católica de Puerto Rico, y profesor asociado en la Facultad de Teología de la Pontificia Universidad Javeriana en Bogotá, Colombia. Ha publicado numerosos artículos de teología, filosofía e Inteligencia Artificial en libros, revistas indexadas y congresos. Desde mayo del 2024 pertenece al Grupo de Trabajo de Frontera Tecnológica del Consejo Episcopal Latinoamericano y del Caribe (CELAM), equipo que redactó un documento sobre inteligencia artificial que fue sometido en febrero del 2025.



**Karina Pedace**: doctora en filosofía por la Universidad de Buenos Aires (UBA), Argentina. Investigadora y profesora adjunta de filosofía contemporánea en la Facultad de Filosofía y Letras de la UBA y del Departamento de Humanidades y Ciencias Sociales en la Universidad Nacional de La Matanza (UNLaM), Argentina. Integra el Instituto de Investigaciones Filosóficas de la Sociedad Argentina de Análisis Filosófico, CONICET. Cofundadora del Grupo de Investigación de Inteligencia Artificial, Filosofía y Tecnología (Grupo GIFT). Coordinadora general de la Red de Mujeres Filósofas de América Latina-UNESCO. Integrante del Comité Asesor de la Redbioética UNESCO. En 2022 fue distinguida internacionalmente entre las *100 Brilliant Women in Artificial Intelligence Ethics* (https://womeninaiethics.org/the-list/of-2022/).

**Tobías J. Schleider**: doctor en filosofía del derecho por la Universidad de Buenos Aires (UBA), Argentina. Abogado y especialista en derecho penal por la Universidad Nacional de Mar del Plata (UNMdP), Argentina. Profesor titular ordinario en la Universidad Nacional del Sur (UNS), Argentina, y profesor de grado y posgrado en la UNMdP e investigador en ambas instituciones y en el Instituto Latinoamericano de Seguridad y Democracia (ILSED). Coordina la licenciatura en seguridad pública de la UNS y dirige el Grupo de Investigación "Seguridad en Democracia". Es, además, miembro de la Comisión Directiva del ILSED. Es autor, editor y traductor de publicaciones especializadas y de divulgación: *Manual de políticas públicas de seguridad ciudadana en Latinoamérica* (Didot, 2024) y *Guía para ciudades más seguras* (CAF, 2021). Interviene regularmente en proyectos de investigación junto a organismos nacionales e internacionales como EACEA, BID, CAF, OEA, FES, Fondo Social Europeo, UCL y ministerios del Estado argentino.

Número especial

Septiembre 2026

REVISTA IBEROAMERICANA DE CIENCIA, TECNOLOGÍA Y SOCIEDAD

OEI

PONTIFICIA COMISIÓN PARA AMÉRICA LATINA

# Compromisos éticos frente a la **inteligencia artificial**



ISSN:1850-0013