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*

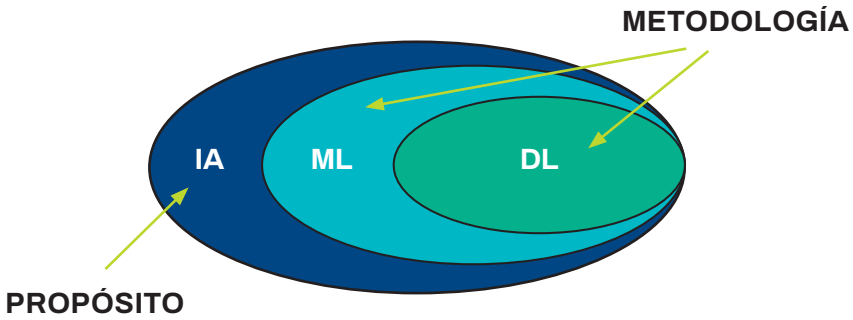
Componente	Función	Descripción
Datos	Qué aprende la IA	Grandes cantidades de información en texto e imagen
Algoritmo	Cómo aprende la IA	Parte programada a mano que determina de qué forma se procesan los datos

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.