DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

Entre sintaxis y normatividad: una aproximación lógica y computacional a los problemas éticos de la inteligencia artificial

Jesús Jasso Méndez*

There is a fallacy often committed in discussion of Artificial Intelligence, especially AI of superhuman capability. Someone says: 'When technology advances far enough, we'll be able to build minds far surpassing human intelligence. Now, it's obvious that how large a cheesecake you can make depends on your intelligence. A superintelligence could build enormous cheesecakes - cheesecakes the size of cities - by golly, the future will be full of giant cheesecakes! The question is whether the superintelligence wants to nbuild giant cheesecakes. The vision leaps directly from capability to actuality,without considering the necessary intermediate of motive.
Eliezar Yudkowsky

Introducción

La relación entre ética e inteligencia artificial (IA) suele analizarse desde distintos campos y enfoques *ex. gr.* filosóficos, sociales, políticos, jurídicos, económicos, tecnológicos. Sin embargo, esta relación puede examinarse

* Profesor-investigador de la Academia de Filosofía e Historia de las Ideas de la Universidad Autónoma de la Ciudad de México, área: Lógica y Epistemología. Profesor en la Universidad Nacional Autónoma de México, área: Lógica y Filosofía de la Mente. Presidente de la Academia Mexicana de Lógica, A. C. (2019-2021). Miembro del Sistema Nacional de Investigadoras e Investigadores (Nivel II). Integrante del Programa de investigación 056 del Colegio de Humanidades y Ciencias Sociales, UACM. Actualmente es miembro del Comité Editorial de *Andamios*. ORCID: 0000-0002-3091-5258 Correos electrónicos: jesus.jasso@uacm.edu.mx; jesusjasso@filos.unam.mx

desde un plano más básico, el que conjuga la lógica y la teoría de la computabilidad. En este nivel, las cuestiones éticas no se plantean exclusivamente en términos de señalar valores o formas correctas de justificar la resolución de problemas teórico-prácticos por parte de un agente artificial, sino a partir de las restricciones inherentes a lo que puede ser formalizado, medido cuantificacionalmente o computado.

¿Es lógicamente posible formalizar de manera completa principios éticos a partir de algún lenguaje lógico altamente expresivo susceptible de implementación computacional?

La teoría de la computabilidad ofrece una respuesta potente pero limitada. Por una parte, algunos sistemas normativos pueden modelarse mediante lenguajes lógicos y ejecutarse algorítmicamente; por otra, los resultados estándar en torno a la indecidibilidad[1] demuestran, para ciertos casos, la inexistencia de un procedimiento algorítmico capaz de incluir al conjunto completo de instancias. En este último escenario se encuentra la normatividad ética.[2] La ética computacional subyacente al desarrollo de las IAs no se limita a las dificultades propias al diseño de sistemas artificiales, sino a cuestiones profundas de limitación lingüística, de limitación teórica.

De acuerdo con las posibilidades del modelaje lógico y computacional, existe un problema general absolutamente no menor relacionado con la potente noción de algoritmo. Si por algoritmo se entiende un procedimiento formal –tipo máquina de Turing: *effective method* (Turing, 1936-1937, p. 230-235)–, su funcionamiento implica la dimensión normativa de seguir pasos ordenados para resolver un problema, para realizar una tarea. En otras palabras, las decisiones automatizadas dependen por definición de un conjunto finito y explícito de reglas. ¿Todo se puede formalizar y computar? La respuesta es no. Existe un conjunto finito de información científica y

[1] En términos generales, la indecibilidad en lógica y teoría de la computabilidad se refiere a casos científicos y matemáticos para los cuales no puede existir algún algoritmo general que siempre dé una salida informacional del tipo sí o no, a partir de un número finito de pasos. Ver por ejemplo, Turing (1936-1937, p. 259-265).

[2] La normatividad ética refiere a las relaciones necesarias entre principios, reglas, criterios, valores, y el diseño, desarrollo y uso de los sistemas inteligentes. En términos generales, aceptar una norma nos compromete con ciertas inferencias y rechaza otras. La dimensión inferencial de la normatividad tiene como finalidad entonces regular acciones, razonamientos y justificaciones en el orden humano y artificial.

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

humanística susceptible a la estructuración lógica y un conjunto de contenidos científicos y humanísticos que, por su naturaleza deliberativa,[3] no forman parte de la cardinalidad lingüística cuantificada *i.e.* tales contenidos son altamente difíciles de codificar al interior de un procedimiento algorítmico. Paradigmáticamente, valoraciones éticas como *bueno, malo, justo, correcto* o *incorrecto* por sus propiedades conceptuales y aplicativas (*ex. gr.* contexto, pluralidad de valores, ambigüedad lingüística, intencionalidad, desacuerdo) no se capturan –al menos no toda su trama y urdimbre– en funciones calculables a lo largo de una estructura computacional que garantice su completa formalización y decibilidad.[4]

Este último aspecto entre dominios formalizables-computables nos conduce naturalmente a cuestiones relacionadas con la *atribución de eticidad a Ias Inteligencias Artificiales*, *i.e.*, a preguntarnos sobre el carácter ético de modelos y algoritmos diseñados no sólo para procesar información y desplegar resultados inferencialmente, sino para simular capacidades similares a las humanas en la toma de decisiones intencionadas, *ex. gr.*, razonar, resolver problemas, comunicar, aprender, justificar resultados. En esta línea, la atribución de moralidad y corrección ética está en juego cuando los sistemas artificiales median las decisiones de los razonadores humanos, resultados que impactan las esferas públicas y privadas de los individuos. Sobre esta posibilidad, consumada actualmente, el compromiso analítico de los especialistas –en lógicas y teorías computacionales, en principio– alrededor del diseño y funcionamiento de inteligencias artificiales, debe dirigirse al diseño de sistemas capaces de salvaguardar, por una parte, la corrección de acciones

[3] La expresión 'naturaleza deliberativa' refiere a la información científica y humanística proveniente de aspectos cognitivos y metodológicos no definitivos *ex. gr.* marcos conceptuales abiertos a constante corrección, tensión entre perspectivas sobre cuestiones compartidas, narrativas científicas basadas en la percepción de los especialistas, acuerdos y desacuerdos modificables entre expertos, contenidos aceptables aunque altamente falibles por su alta complejidad, y valores científicos y humanísticos no uniformes.

[4] La expresión "completa formalización y decibilidad" en términos generales refiere, por una parte, a un dominio *ex. gr.* de enunciados e incluso de reglas expresados íntegramente dentro de un sistema lógico-matemático; la decibilidad, por su parte, consiste en un procedimiento algorítmico capaz de resolver en cualquier caso correctamente a una pregunta (entrada- proceso y ejecución finita-respuesta, en términos de sí o no para cualquier entrada). Respecto a la noción de formalización se puede consultar *ex. gr.* Hulth y Ryan (2004); respecto a la noción de decibilidad se puede consultar *ex. gr.* Sipser (2013, p. 193-210).

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

y prácticas sociales como resultado de entradas informacionales ordenadas inferencialmente; por otra, la justificación de resultados de acuerdo con éticas humanas bien construidas y valoradas públicamente.

El presente artículo examina la relación entre Ética e Inteligencia Artificial (IA) restringiendo el análisis al ámbito de la lógica y la teoría de la computabilidad. A partir de los problemas contemporáneos de responsabilidad y agencia, sesgo algorítmico, opacidad, autonomía y alineación de valores, se argumenta que dichos problemas surgen de tensiones conceptuales internas a las nociones de formalización y modelación algorítmica. A través de ex. gr. Alan Turing (1936), Andreas Matthias (2004), Vikas Hassija (2004), Joy Buolamwini (2018), Scott Aaronson (2013), se sostiene que los sistemas algorítmicos operan en un nivel sintáctico que no necesariamente agota la dimensión semántica ni inferencial de la normatividad ética. En consecuencia, los problemas éticos de la IA deben entenderse como expresiones de los límites estructurales de la formalización y la decibilidad algorítmica. Este artículo finalmente, propone brevemente una reconstrucción del enfoque pragmático-inferencial, inspirado en Luciano Floridi *et al.*, (2018); donde el componente de la normatividad se sitúa constructivamente en el diseño y funcionamiento de los sistemas inteligentes.Ética, inteligencia artificial y problemas ético-técnicos: ¿es posible atribuir eticidad a los sistemas artificiales?

Responsabilidad y agencia

El problema de la *responsabilidad y agencia* surge de la tensión entre la autonomía funcional de las IAs y los mecanismos que los razonadores humanos utilizamos para atribuirles propiedades éticas. ¿Quién es responsable de aquellas decisiones con efecto público, si éstas ocurren por mediación de sistemas algorítmicos inteligentes? La atribución de responsabilidad de acuerdo con este desajuste implica "facing a responsibility gap, which cannot be bridged by traditional concepts of responsibility ascription"[5] (Matthias, 2004, p. 175). Luego, la adjudicación ética en torno a la responsabilidad será, por lo menos, imprecisa y altamente polémica: *responsibility gap*.[6]

[5] "enfrentarse a un vacío de responsabilidad que no puede subsanarse con los conceptos tradicionales de atribución de responsabilidad" (La traducción es responsabilidad mía).

[6] Es interesante observar que la noción de *responsability gap* implica una contra-intuición de

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

En esta línea, si las IAs fuesen agentes morales, entonces funcionarían de acuerdo con razones y en función de consideraciones normativas que justifiquen su operación y despliegue de información. Sin embargo, claramente los sistemas artificiales son interesantes en la medida en que son capaces de procesar información y ejecutar tareas mediante instrucciones específicas. Pese a ello, esta capacidad no es simétrica con la facultad de comprender lineamientos normativos, *i.e.*, interpretar, justificar y aplicar principios éticos de corrección al momento de resolver problemas informacionales. Luego, el núcleo del problema ético, en este caso, consiste en enfrentar dos intuiciones contrarias, por una parte, las consecuencias informacionales desde las IAs tienen no autocontradictoriamente un efecto moralmente relevante y, por otra parte, las IAs por sí mismas no son auténticas agencias morales.

Pensemos en un caso estándar. Un carro autónomo (sin chofer) va por la carretera. Un venado cruza la autopista de manera inesperada. El sistema del carro debe decidir en una milésima de segundo qué hacer. El carro hace una maniobra que termina causando daño al venado (ver, *ex. gr.*, Nyholm y Smids, 2016). En este caso, ¿quién es el responsable del accidente?, ¿a quién atribuimos responsabilidad, al programador, a la empresa, al usuario ausente, al organismo de tránsito, al sistema artificial? Esta situación constituye un escenario donde el resultado es moralmente evaluable, pero el origen intencional no es claramente atribuible, no al menos, a un único agente. Casos simétricos con impacto individual e institucional ocurren actualmente, no obstante, como se ha señalado, la adjudicación es difusa. La

atribución, la cual sitúa perfectamente el problema constructivo de la responsabilidad-agencia. En palabras de Andreas Mattias (2004): "Autonomous, learning machines [...] create a new situation, where the manufacturer/operator of the machine is in principle not capable of predicting the future machine behaviour any more, and thus cannot be held morally responsible or liable for it." ("Las máquinas autónomas y con capacidad de aprendizaje [...] dan lugar a una nueva situación en la que el fabricante u operador de la máquina, en principio, ya no es capaz de predecir el comportamiento futuro de ésta y, por lo tanto, no se le puede considerar moralmente responsable o culpable de ello". La traducción es responsabilidad mía.) En este caso, por una parte, no es posible atribuir responsabilidad moral al razonador humano y sí, al parecer, al sistema artificial. Sin embargo, las IAs no comprenden marcos normativos, tal que, no constituyen genuinos agentes morales. Luego, hay generadores artificiales de circunstancias morales y, al tiempo, tales generadores no son sujetos morales. Existe una zona gris de atribución. La atribución moral no se da de manera directa a los razonadores humanos, ni a los razonadores artificiales. Y tampoco a los usuarios.

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

atribución de moralidad y justificación ética a las IAs requiere criterios más comprehensivos que la mera capacidad de procesar y desplegar información como fuente de toma de decisiones.

Finalmente, en función con la Teoría de la computabilidad, los sistemas algorítmicos tienen limitaciones estructurales no sólo para decidir qué hacer o qué resultado generar, sino para estructurarse como procedimientos completos y finitos que permitan trazar una cadena de inferencias recursivas para realizar tareas en cualquier caso –indecibilidad–. De acuerdo con Scott Aaonson (2013), existen problemas bien definidos que ningún algoritmo puede resolver por poderoso que sea el sistema de cómputo (cfr. Aaronson, 2013, p. 18-28). Este escenario fortalece la idea sobre la imposibilidad de tratar a los sistemas inteligentes como agentes morales genuinos, al estar constructivamente restringidos a condiciones formales cuyos límites impiden considerar de manera equivalente las capacidades algorítmicas y las agencias humanas.

Sesgo algorítmico

El problema del *sesgo algorítmico* está estrechamente relacionado con el problema de *responsabilidad* y *opacidad*.[7] En términos analíticos es posible distinguir entre el carácter matemático estructural de las IAs y los efectos sociales que pueden surgir de su operación. Mediante una apariencia de neutralidad computacional, los sistemas inteligentes –*ex. gr.*, sistemas de reclutamiento, sistemas de reconocimiento facial, sistemas de evaluación de riesgos, sistemas para la procuración e impartición de justicia, sistemas de predicción– en un contexto real, pueden promover, fortalecer, reproducir, legitimar, situaciones de inequidad y exclusión social.

Los algoritmos computacionales no surgen por sí mismos. El procesamiento informacional artificial es consecuencia de un trabajo conceptual e intencional de los razonadores humanos. A partir del diseño, entrenamiento, ejecución y despliegue de información, las IAs se incorporan a un entramado socio-cultural de corte normativo. Constructivamente, en términos generales, los sistemas inteligentes se basan en recortes de información, *i.e.*, datos sesgados, decisiones intencionales de los programadores, bases de in-

[7] Más adelante se abordará el problema de la *opacidad* y su extensión.

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

clusión social no equitativa, interpretaciones automatizadas simplificadoras de problemas (*ex. gr.* sociales altamente complejos). Estas circunstancias hacen que el mecanismo completo de una IA sea normativamente interesante, al producir resultados informacionales necesaria y sistemáticamente desfavorables para un conjunto de individuos o un conjunto de grupos de individuos. El problema del sesgo algorítmico refiere a estas condiciones.

Al respecto, es posible encontrar líneas de investigación con amplio reconocimiento, *ex. gr.*, Buolamwini y Gebru (2018): sesgo en reconocimiento facial; Conitzer *et al.* (2023): sesgo en algoritmos de contratación. De acuerdo con el primer caso, "...machine learning algorithms can discriminate based on classes like race and gender"[8] (Buolamwini y Gebru, 2018). De acuerdo con el segundo caso, sistemas inteligentes de algunas empresas, al estar entrenados con datos históricos dominados por hombres, devaluaban automáticamente solicitudes que incluían indicadores asociados a mujeres:

> IF YOU READ news articles on the ethics of AI, you will repeatedly see the phrase "algorithmic bias" popping up. It refers to algorithms producing results that appear racist or sexist or displaying other forms of unfair bias. For example, when Amazon built a machine-learning model to score job applications, trained on historical hiring data at the company, it discovered the system downgraded female applicants-or anyone who mentioned activities associated with women (coach of women's soccer league, for example). (Conitzer *et al.*, 2023)[9]

Estas circunstancias y sus implicaciones muestran que el sesgo algorítmico no es una condición limitada al funcionamiento inferencial de los siste-

[8] "los algoritmos de aprendizaje automático pueden discriminar en función de categorías como la raza y el género". (La traducción es responsabilidad mía).

[9] "SI LEES artículos de prensa sobre la ética de la IA, verás aparecer una y otra vez la expresión "sesgo algorítmico". Se refiere a los algoritmos que producen resultados que parecen racistas o sexistas, o que muestran otras formas de sesgo injusto. Por ejemplo, cuando Amazon creó un modelo de aprendizaje automático para evaluar solicitudes de empleo, entrenado con datos históricos de contratación de la empresa, descubrió que el sistema penalizaba a las candidatas mujeres, o a cualquier persona que mencionara actividades asociadas con las mujeres (entrenadora de una liga de fútbol femenino, por ejemplo)". La traducción es responsabilidad mía.

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

mas artificiales, sino una propiedad de las IAs, la cual produce evidentes tensiones estructurales más profundas entre el desarrollo de sistemas computacionales automatizados y aspectos constructivos de los valores éticos socialmente dominantes.

¿Quién es responsable cuando una IA produce decisiones no justas? Resulta fácil inferir la relación entre el problema de la responsabilidad y agencia, y el sesgo. A diferencia de la resolución humana de problemas prácticos y conceptuales, en cuyo caso la atribución de responsabilidad tiene por lo general trayectoria definida hacia individuos o instituciones específicas, las IAs, como se ha señalado, distribuyen la agencia sin trayectoria clara a múltiples actores, *ex. gr.*, programadores, diseñadores, patrocinadores, empresas, usuarios. Esto hace que, entre mayor sea la autonomía funcional artificial,[10] la noción de responsabilidad estándar se diluya. Luego, cuando un algoritmo discrimina, no es fácil establecer qué agente es moral o legalmente responsable. El problema de la responsabilidad y el sesgo resultan interesantes particularmente por dos aspectos: i. permiten identificar una tendencia sistemática de los sistemas inteligentes a favorecer o perjudicar a un grupo de individuos o instituciones específicas; ii. nos permiten reconocer propiedades inherentes al funcionamiento algorítmico de los sistemas inteligentes. Los algoritmos no eliminan valores humanos, los reproducen sesgadamente reforzando marcos específicos con referencia moral.

Opacidad y autonomía

El problema de la opacidad (*black box*)[11] desde una perspectiva lógica y computacional es muy interesante. Se podría pensar, *ex. gr.*, que los diseñadores de sistemas computacionales inteligentes tienen acceso claro y completo a su funcionamiento. Sin embargo, los sistemas automáticos son altamente complejos en términos de su estructura, operación y alcances para la resolu-

[10] Más adelante se abordará el problema de la *autonomía* y su extensión.

[11] El término *black box* es comúnmente utilizado para referirse al problema de la opacidad de los sistemas inteligentes. Uno de los artículos con mayor influencia sobre este punto y ampliamente recomendado para analizar conceptual y aplicativamente este problema es Hassija *et al.* (2024), aunque existen en la literatura muchos otros.

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

ción de problemas. Esta inaccesibilidad epistémica constituye el centro del problema de la opacidad:

> Recent years have seen a tremendous growth in Artificial Intelligence (AI)-based methodological development in a broad range of domains. In this rapidly evolving field, large number of methods are being reported using machine learning (ML) and Deep Learning (DL) models. Majority of these models are inherently complex and lacks explanations of the decision making process causing these models to be termed as 'Black-Box'. One of the major bottlenecks to adopt such models... is the difficulty in interpreting them. (Hasijja et. al, 2024, p. 46)[12]

La inaccesibilidad epistémica al completo funcionamiento de un sistema inteligente produce problemas en el nivel de la normatividad de las IAs. Si no es posible tener acceso cognitivo íntegro a cómo una IA opera y despliega información, entonces principios éticos fundamentales, *ex. gr.*, la responsabilidad, la justicia, la rendición de cuentas en la toma de decisiones podrían claramente trastocarse. ¿Cómo podemos evaluar decisiones artificiales moralmente teniendo en mente la inaccesibilidad cognitiva?

La opacidad o *black box* tiene consecuencias anómalas importantes a nivel ético y a nivel epistemológico. En primer lugar, si existen contextos de decisión humanos, artificiales o mixtos donde no se comprende a profundidad lo que está ocurriendo constructivamente al interior del *medio* de decisión, entonces la atribución de responsabilidades sobre el funcionamiento y despliegue de información de un sistema artificial constituye una medida borrosa.[13] ¿Quién hace qué y cómo lo hace de acuerdo con un sistema opaco?

[12] "En los últimos años, se ha producido un enorme crecimiento en el desarrollo de metodologías basadas en la inteligencia artificial (IA) en una amplia variedad de campos. En este ámbito en rápida evolución, se están publicando numerosos métodos que utilizan modelos de aprendizaje automático (ML) y aprendizaje profundo (DL). La mayoría de estos modelos son inherentemente complejos y carecen de explicaciones sobre el proceso de toma de decisiones, lo que ha llevado a que se les denomine "cajas negras". Uno de los principales obstáculos para la adopción de dichos modelos... es la dificultad para interpretarlos". La traducción es responsabilidad mía.

[13] En este caso, la expresión "borrosa" refiere a un contexto donde no es posible concentrarse en un sólo aspecto.

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

Las decisiones públicas y privadas se pueden evaluar integralmente sólo en estados de transparencia y condiciones de explicabilidad. Los objetos de medida son intenciones, objetivos y resultados; al margen de éstos, la atribución del carácter ético de las acciones se limita o incluso queda obstaculizada. La atribución de responsabilidad particularmente a sistemas artificiales opacos queda entorpecida por sus condiciones de posibilidad estructural, algorítmica y computacional.

Por otra parte, los términos de inaccesibilidad epistémica a sistemas inteligentes son un fenómeno contemporáneo en el campo de la epistemología computacional. En términos generales, cualquier planteamiento fundacionista del conocimiento rehúye posiciones escépticas en torno a nuestras explicaciones del mundo (Dancy, 1993, p. 71-104). En el campo de las IAs ocurre lo mismo. La opacidad en este caso es especialmente preocupante.

Pensemos, por ejemplo, en la atribución de la existencia de otras mentes humanas (un problema filosóficamente complicado de solventar). Este conocimiento claramente no se sigue de un conocimiento directo i.e no percibimos otras mentes ni existe una accesibilidad directa al contenido mental de otros razonadores. Sin embargo, existen medios por los cuales la afirmación de los contenidos mentales en tercera persona puede admitirse, *ex. gr.*, condiciones pragmáticas, afirmaciones sobre creencias primitivas en primera persona, mediciones científicas, conocimiento por referencia, sistemas de representación lingüística, atribuciones inferenciales. No obstante, el problema de la opacidad computacional por ahora no tiene opciones definitivas de solución. En este caso, tampoco se tiene conocimiento directo del procesamiento computacional y, adicionalmente, es particularmente difícil atribuir con verdad a un razonador humano conocimiento integral de las condiciones computacionales del sistema artificial. Un programador o diseñador computacional tiene conocimiento importante del sistema, pero el conocimiento situado en cualquier caso de qué hace el sistema y cómo llega exactamente a una salida informacional, es siempre parcial. Actualmente se busca subsanar el problema.[14]

[14] Particularmente, un programador o diseñador de un sistema no sabe en su totalidad, *ex. gr.*, la combinatoria entre millones de variables internas del sistema, cuáles son los patrones emergentes que el sistema de datos aprende, cómo el sistema funcionará en circunstancias no consideradas, por qué un sistema tomó una decisión específica en un caso particular. En-

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

En un contexto donde las decisiones humanas se encuentran mediadas por sistemas inteligentes, ¿qué agente o agentes deciden realmente y bajo qué condiciones puede considerarse que tales decisiones surgen de condiciones, libres, responsables y justificables? Esta pregunta nos conduce naturalmente al problema de la *autonomía*[15] de sistemas inteligentes.

Los problemas de la *opacidad* y *autonomía* se encuentran estrechamente relacionados. Si en algún campo las decisiones humanas dependen de la intervención algorítmica y, sobre esta intervención no se tiene un acceso cognitivo completo y claro, mínimamente se infieren dos circunstancias indeseables: I. delegar la toma de decisiones con impacto social a sistemas artificiales deflaciona la autonomía humana; II. aceptar la autonomía de sistemas artificiales opacos implica atribuir autodeterminación a sistemas inteligentes aun cuando no sea posible comprender totalmente su funcionamiento.

De acuerdo con el punto II, ¿en qué sentido es posible atribuir autonomía a las IAs? En principio, se debe distinguir entre dos maneras interesantes de conceptualizar la autonomía: autonomía racional y autonomía funcional. El primer caso refiere a la capacidad de un agente para darse a sí mismo sus propias normas y asumir la responsabilidad de sus decisiones conforme a principios reflexivos. El segundo caso refiere a la capacidad de un agente para actuar sin supervisión constante. En el campo de la computabilidad se concibe a los sistemas inteligentes como estructuras algorítmicas *automatizadas*, *i. e.*, sistemas que pueden funcionar independientemente del programador una vez diseñados y ejecutada su capacidad de ordenamiento informacional. Este tipo de autonomía no implica por parte del agente artificial, deliberación racional, responsabilidad moral, tampoco justificación en la toma de decisiones (aspectos constructivos del primer tipo de autonomía). En este sentido, la autonomía artificial es el ejemplo paradigmático de una autonomía funcional, autonomía alejada de una auto-organización racional.

frentar el problema de la opacidad o la *black box* representa uno de los retos más importantes en los campos de aprendizaje automático y aprendizaje profundo. Ver, *ex. gr.*, Tripathy y Seetha (2024), Mehta, *et al.* (2023).

15 La autonomía se entiende como la capacidad de un agente por autolegislarse y hacerse responsable de sus decisiones bajo un marco general de racionalidad. Esta concepción se remonta al enfoque kantiano donde se concibe la autonomía como la capacidad de la voluntad de un agente para darse a sí misma la ley moral. Este aspecto de darse a sí mismo la ley moral implica para Kant (1980) actuar racionalmente y ser responsable del ejercicio de la libertad.

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

El núcleo de la tensión conceptual respecto a la opacidad en relación con la autonomía es de naturaleza ética. De acuerdo con el concepto de autonomía desde una vertiente kantiana, las IAs no serían autónomas, lo cual sin más sería intuitivamente aceptable; pero si se emplea un sentido funcional de autonomía, las IAs serían autónomas, aunque sin responsabilidad moral (sistemas no deliberativos). Esta última circunstancia parecería librar a las IAs de todo compromiso normativo, lo cual sería inaceptable ya que el carácter funcional del sistema artificial conlleva decisiones con impacto en autonomías racionales. Luego, las IAs no tienen responsabilidad moral y, al tiempo, deben evaluarse moralmente. Este resultado nos conduce a lidiar con un concepto de autonomía propio al campo de la computabilidad conceptualmente inconsistente y necesariamente, en términos pragmáticos, evaluable.

Alineación de valores (Value alignment problem)

¿Es posible codificar normatividad ética? El problema de alineación de valores refiere a la dificultad que enfrentan los diseñadores, programadores, empresas, patrocinadores y usuarios, para garantizar un procesamiento informacional inteligente de acuerdo con marcos éticos bien valorados en el campo de los razonadores humanos.

El problema no es trivial. Las IAs toman decisiones autónomas funcionales en distintos ámbitos de acción humana, *ex. gr.*, los sistemas de salud, seguridad pública, acción empresarial, administración pública, administración de justicia, sistemas de reclutamiento. La preocupación ética se hace evidente al preguntar sobre el alcance de justicia existente en el proceso decisorio. ¿La programación de la IAs puede efectuarse con neutralidad informacional?

Las decisiones humanas mediadas por sistemas artificiales priorizan intenciones u objetivos mediante el despliegue informacional. Estas intenciones u objetivos –*ex. gr.*, de los gobiernos, empresas, instituciones patrocinadoras– pueden ser altamente incompatibles con los derechos fundamentales de los seres humanos, provocando con ello condiciones sociales altamente indeseables. Alinear las IAs implica al menos promover en los desarrolladores dos exigencias. Por una parte, desarrollar sistemas con lenguaje más expresivos, más inteligentes *i. e.* con mayor capacidad de procesamiento informacional alterno de distintos campos. Por otra, desarrollar

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

agencias funcionales inteligentes consistentes con la posibilidad de generar resultados automatizados dependientes de los intereses humanos más justos y racionales. Aunque la tarea parezca casi imposible se trata de enfrentar "the impossibility of defining true human purposes correctly and completely [... although] what humans try to do whereby humans attempt to imbue machines with their own purposes-is destined to fail" (Russell, 2019, p. 137).[16]

Si la autonomía de las IAs es de tipo funcional no constructiva de una agencia moral genuina, la alineación parece ser una tarea no delegable a los sistemas y su automatización. En todo caso, la responsabilidad de esta empresa recae no sólo en los diseñadores de sistemas inteligentes sino en aquellos que ejercen una normatividad humana singular al apoyar la codificación de objetivos individuales, sociales, corporativos, educativos, tecnológicos, políticos; implicados en el desarrollo social de la agencia humana. Esto es altamente interesante. Un sistema automatizado puede ser altamente eficiente y resultar altamente problemático cuando su preparación operativa parte de un registro histórico de datos reforzadores de injusticias sociales (ver Buolamwini, 2018).

En suma, desde una perspectiva lógica y computacional los problemas asociados con la relación entre ética e inteligencia artificial –responsabilidad y agencia, sesgo algorítmico, opacidad (*black box*) y alineación de valores (*value alignment problem*)– requieren mecanismos analíticos interdisciplinarios que integren una evaluación ética de las ciencias computacionales y sus aplicaciones en los distintos campos de interacción humana, donde las IAs sean mediadoras para la toma de decisiones. Particularmente, se requiere un *human oversight*[17] en torno al diseño de modelos computacionales automatizados, con la finalidad de preservar las garantías fundamentales.

16 "la imposibilidad de definir de forma correcta y completa los verdaderos fines del ser humano... [aunque] lo que los seres humanos intentan hacer –es decir, dotar a las máquinas de sus propios fines– está destinado al fracaso". La traducción es responsabilidad mía.

17 La expresión *human oversight* refiere a un principio ético y legal el cual prescribe la supervisión humana sobre los sistemas de inteligencia artificial para asegurar que su funcionamiento, *ex. gr.*, decisiones, acciones o despliegues de información se encuentren bajo el control e intervención de los razonadores humanos.

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

Ética e inteligencia artificial. Un enfoque pragmático-inferencial desde Luciano Floridi (2018)

Los problemas inherentes a la relación entre ética y las inteligencias artificiales trascienden la pregunta estándar: ¿las máquinas piensan? Las consideraciones filosóficas, lógicas y computacionales deben centrar su atención en dos aspectos: i. analizar bajo qué criterios normativos generales la agenda de investigación en sistemas artificiales debe fundarse; y ii. identificar desde qué marco normativo pueden justificarse los resultados generados por sistemas capaces de seguir patrones y funcionar autónomamente.

La perspectiva pragmática-inferencial aborda los puntos I y II a partir de una consideración regulativa: el análisis ético sobre las IAs no consiste en establecer relaciones entre marcos normativos y sistemas inteligentes en términos de dos campos de investigación independientes, donde el primer campo se aplica al segundo para analizarlo y regularlo. En cualquier caso, la relación correcta entre normatividad e IAs implica una perspectiva constructiva de los sistemas computacionales donde el componente ético constituye una propiedad inherente a tales sistemas. Las IAs deben funcionar como estructuras generadoras de información para la toma de decisiones en términos de una *justificación pública* orientada a regular *consecuencias informacionales*. No toda consecuencia informacional podría ser socialmente admisible, aun cuando sea inferencialmente posible.

Las IAs son interesantes no sólo por el modo único de procesar grandes *clusters* de información, sino porque en la actualidad representan, por una parte, vías de conocimiento directo de distintas realidades[18] y, por otra, dispositivos artificiales creadores de nuevas formas de existencia. En este contexto, la racionalidad algorítmica implica cadenas de inferencias sintácticas que codifican distintos intereses humanos de diferente orden, concentrados en amplios grupos, *ex. gr.*, metafísicos (realidad), epistemológicos (conoci-

[18] Asumo que las IAs artificiales actualmente se han convertido en un medio sensorial de los razonadores humanos al nivel de nuestros sentidos estándar *i. e.* vista, tacto, oído, gusto, olfato. En este sentido, los sistemas artificiales son fuente de conocimiento directo y no conocimiento mediado, como podría pensarse. La diferencia entre este nuevo sentido artificial es que adicionalmente está cargado de teoría. Estas consideraciones serán centrales en otro artículo.

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

miento), axiológicos (supuestos normativos). La agenda de investigación en IA debe incorporar un propósito fundacional: relacionar la información de los distintos grupos, de tal suerte que pueda alcanzar la creación y desarrollo de una *Good AI Society*. Tomo la expresión *Good AI Society* de Floridi, *et al.* (2018) para referir a una sociedad con una IA humanamente responsable:

> AI is not another utility that needs to be regulated once it is mature. It is a powerful force, a new form of smart agency, which is already reshaping our lives, our interactions, and our environments. AI4People was set up to help steer this powerful force towards the good of society, everyone in it, and the environments we share...to propose a series of recommendations for the development of a Good AI Society. (Floridi *et al.*, 2018)[19]

En esta línea, el enfoque pragmático-inferencial que propongo debe desarrollarse en términos de una reconfiguración institucional del razonamiento algorítmico. Las decisiones automatizadas deben integrar aspectos éticos fundamentales que regulen inteligencias humanas y artificiales. Los procesos automatizados de decisión deben partir de diseños informacionales y desplegar resultados consistentes con principios de bienestar social, justicia, autonomía y explicabilidad. Estas condiciones juntas serán un contrapeso a las condiciones reales de inaccesibilidad epistémica. Aun cuando los razonadores humanos no tengan conocimiento completo del funcionamiento de las IAs, las condiciones estructurales de los sistemas inteligentes disminuyen considerablemente la posibilidad de inferir resultados éticamente indeseables. En una línea simétrica Florindi señala:

> If we are increasingly going to use the assistance of or delegate decisions to AIs, we need to make sure these systems are fair in their

[19] "La IA no es un servicio más que deba regularse una vez que haya alcanzado su madurez. Es una fuerza poderosa, una nueva forma de agencia inteligente, que ya está transformando nuestras vidas, nuestras interacciones y nuestros entornos. AI4People se creó para ayudar a orientar esta poderosa fuerza hacia el bien de la sociedad, de todas las personas que la integran y de los entornos que compartimos... con el fin de proponer una serie de recomendaciones para el desarrollo de una sociedad con una IA responsable". La traducción es responsabilidad mía.

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

> impact on people's lives, that they are in line with values that should not be compromised and able to act accordingly, and that suitable accountability processes can ensure this. (Floridi *et al.*, 2024, p. 1871)[20]

La perspectiva pragmática-inferencial de las IAs desplaza el debate de la mera regulación normativa de los sistemas artificiales, a la posibilidad constructiva de un ecosistema ético (humano-artificial) donde los sistemas computacionales autónomos se integran a la lógica social, *i. e.*, al desarrollo de las funciones humanas públicas y privadas sin comprometer la dignidad de individuos o grupos de ellos.

Esta perspectiva contempla los problemas revisados en la sección anterior. Si existe ecosistema ético, la inaccesibilidad epistémica no generará una zona gris de imputación de responsabilidad, esta trazabilidad estará en gran proporción medida desde un inicio en las condiciones estructurales. De acuerdo con el problema de la autonomía, si el diseño y funcionamiento de los sistemas artificiales implica y despliega valores sociales altamente reconocidos, la autonomía funcional de las IAs y la autonomía racional humana estarán estrechamente vinculadas. La operación algorítmica constituirá una extensión institucional de criterios normativos humanos ya presentes en la selección de información, la estructura algorítmica, los criterios de optimización de máquinas inteligentes y los objetivos de despliegue informacional para la toma de decisiones.[21]

Las IAs desde una perspectiva pragmática-funcional no se conciben únicamente como herramientas de trabajo, investigación o entretenimiento, sino como dispositivos capaces de generar conocimiento directo y por

[20] Esta observación aparece en Novelli *et al.* (2024). La traducción es responsabilidad mía: "Si cada vez vamos a recurrir más a la ayuda de la inteligencia artificial o a delegar decisiones en ella, debemos asegurarnos de que estos sistemas tengan un impacto justo en la vida de las personas, que se encuentren en consonancia con los valores que no deben verse comprometidos y que sean capaces de actuar en consecuencia, y que existan procesos de rendición de cuentas adecuados que lo garanticen".

[21] En este caso, si los sistemas inteligentes infirieran información anómala o inesperada, esta información se evaluaría de acuerdo con un marco de referencia normativo diseñado de acuerdo con las condiciones de una Good IA Society al interior de un ecosistema humano-artificial ético.

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

referencia[22] con la capacidad de crear realidades y transformar entornos de organización social. Si bien, en estos casos podría surgir la tensión entre una tendencia a delegar decisiones con propiedades éticas en algoritmos y la perspectiva de rechazar agendas de investigación en IAs por considerarse sistemas dehumanizados, la perspectiva pragmatista-inferencial evita esta tensión. Las IAs estructuralmente se conciben subordinadas a condiciones éticas intrínsecas a su configuración, además de incorporar *variables normativas de revisión humana constante*, *ex. gr.*, auditorías tecnológicas de carácter ético y evaluación contextual/situada.

El problema del sesgo algorítmico también bajo esta perspectiva es posible deflacionarse. Si se regulan las condiciones de inferencia pública, un algoritmo codificado pragmática-inferencialmente puede contribuir a la asignación justa de oportunidades, a la distribución-gestión justa de recursos socio-económicos, al establecimiento de mecanismos institucionales de carácter público que amparen los derechos humanos de las poblaciones. En esta línea, el problema de la opacidad-explicabilidad puede enfrentarse al explicitar las razones prácticas que justifican desde un inicio el diseño e implementación sistemas artificiales con propiedades éticas.

Por su parte, el problema de la alineación desde el enfoque pragmático-inferencial puede solucionarse al estar obligados modalmente a la codificación de valores y objetivos humanos altamente aceptables y extendidos en los sistemas artificiales. Esta codificación deberá estar en la base del diseño y operación de las IAs. En cualquier caso, el problema de la alineación artificial exigirá una solución al conflicto humano en torno a un consenso general sobre qué valores deben priorizarse en las diferentes esferas de acción humana y artificial. Este aspecto, sin duda, refuerza la relevancia de un enfoque pragmatista-inferencial interesado en la constitución de ecosistemas éticos como sistemas que exigen procesos permanentes de evaluación crítica del desarrollo de sistemas inteligentes en conjunción con las instituciones sociales.

Finalmente, una aproximación analítica sobre la ética de la inteligencia artificial implica no perder de vista la no neutralidad tecnológica. Los sistemas artificiales integran procesos inferenciales que determinan sociedades

[22] La expresión "conocimiento por referencia" apunta a un conocimiento representacional basado en descripciones lingüísticas y descripciones proposicionales.

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

y relaciones de poder entre los agentes sociales. El aspecto ético central en la relación entre ética y las inteligencias artificiales será diseñar sistemas eficientes que sean constructivamente consistentes con formas de racionalidad computacional que, a su vez, sean compatibles inherentemente con valores democráticos y pluralistas. De acuerdo con la propuesta pragmática-inferencial, la responsabilidad requiere explicabilidad inferencial, la autonomía refiere a delegación auditada, la alineación constituye un criterio dirigido a la negociación normativa permanente entre los sistemas humano-artificiales. El alcance de estas condiciones hace de las inteligencias artificiales dispositivos algorítmicos expansores de la dimensión humana contemporánea.

Agradecimiento

Agradezco profundamente a la Dra. Elizabeth Jasso Méndez por su lectura atenta de este artículo. Las generosas observaciones de la Dra. Jasso permitieron mejorar la versión final de esta propuesta.

Fuentes consultadas

Aaronson, S. (2013). *Quantum Computing since Democritus*. Cambridge University Press.

Buolamwini, J. y Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. En *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT* 2018), Proceedings of Machine Learning Research*. Núm. 81. pp. 77-91.

Conitzer, V., Hadfield, G. y Vallor, S. (2023). Technical Perspective: the Impact of Auditing for Algorithmic Bias. En *Communications of the ACM*. Vol. 66. Núm. 1. pp. 100-101 DOI: https://doi.org/10.1145/3571152

Dancy, J. (1993). *Introducción a la epistemología contemporánea*. Tecnos. pp. 71-104.

Floridi, L., Cowls, J. y Beltrametti, M. *et al.* (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks,

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

Principles, and Recommendations. En *Minds & Machines*. Núm. 8. pp. 689-707. DOI: https://doi.org/10.1007/s11023-018-9482-5

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M. y Hussain, A. (2024). Interpreting Black-Box Models: a Review on Explainable Artificial Intelligence. En *Cognitive Computation*. Núm. 16. pp. 45-74. DOI: https://doi.org/10.1007/s12559-023-10179-8

Hulth, M. y Ryan, M. (2004). *Logic in Computer Science: Modelling and Reasoning about Systems*. Cambridge University Press.

Kant, I. (1980). *Fundamentación de la metafísica de las costumbres*. Espasa-Calpe.

Matthias, A. (2004). The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. En *Ethics and Information Technology*. Vol. 6. Núm. 3. pp. 175-183. DOI: https://doi.org/10.1007/s10676-004-3422-1

Mehta, M., Palade, V. y Chatterjee, I. (Eds.). (2023). *Explainable AI: Foundations, Methodologies and Applications*. Springer Cham. DOI: https://doi.org/10.1007/978-3-031-12807-3

Nyholm y Smids (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?. En *Ethic Theory Moral Prac*. Núm. 19. pp. 1275-1289. DOI: https://doi.org/10.1007/s10677-016-9745-2

Novelli, C., Taddeo, M. y Floridi, L. (2024). Accountability in Artificial Intelligence: What it is and How it Works. En *AI & Soc*. Núm. 39. pp 1871-1882. DOI: https://doi.org/10.1007/s00146-023-01635-y

Sipser, M. (2013). *Introduction to the Theory of Computation*. Cengage Learning.

Turing, A. (1936-1937). On Computable Numbers, with an Application to the Entscheidungsproblem. En *Proceedings of the London Mathematical Society*. Serie 2. Vol. 42. Parte. I. pp. 230-265.

Tripathy, B. y Seetha, H. (Eds.). (2024). *Explainable, Interpretable, and Transparent AI Systems*, CRC Press. DOI: https://doi.org/10.1201/9781003442509

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349

Yudkowsky, E. (2008). Artificial Intelligence as Positive and Negative Factor in Global Risk. En N. Bostrom y M. Cirkovic (Eds.). (2008). *Global Catastrophic Risks*. Oxford University Press.

DOI: https://doi.org/10.29092/uacm.v23iEspecial.1349