CÓMO CITAR
Lopez, R. F. (2025). *Ética de la inteligencia artificial* de Luciano Floridi. *Ethika+*, (11), 239-251. https://doi.org/10.5354/2452-6037.2025.78105

Ética de la inteligencia artificial de Luciano Floridi

Raúl Fernando López[1]
Estudiante de Licenciatura en Filosofía, Universidad de Chile
raul.lopez.r@ug.uchile.cl

Barcelona: Herder, 2024
464 pp.
ISBN: 978-84-254-5065-5

Sin duda la tecnología es parte cada vez más importante de nuestras vidas, tanto es así que la frontera divisoria entre lo artificial y lo natural parece difuminarse hasta desaparecer. Y es que, según piensan algunos (Hui, 2022), la inteligencia artificial (IA) está abriendo una brecha en el marco conceptual meramente mecánico al que parecía estar confinada en la teoría cartesiana (González, 2016). Pero, ¿es realmente la IA lo que se abre paso en los terrenos metafísicos de la existencia a través de la posibilidad de tener estados mentales, o es que somos nosotros quienes estamos transformando

[1] https://orcid.org/0009-0000-6904-187X

el mundo que nos rodea —y de paso a nosotros mismos— en un entorno digital favorable para el despliegue idóneo de las nuevas tecnologías inteligentes?

Luciano Floridi, considerado una de las mayores autoridades en materia de filosofía y ética de la información, presenta un exhaustivo estudio sobre las nuevas tecnologías de IA desde una perspectiva filosófica, cuyas implicaciones éticas exigen una reflexión profunda. El desarrollo de una investigación de este tipo sin duda requiere de un andamiaje conceptual específico y técnico que permita describir con precisión la problemática abordada. En este sentido, el autor italiano introduce conceptos clave como *infoesfera*, *agencia artificial* y *envoltura*, con el propósito de entender el contexto en el que resulta necesario el trazado de principios y prácticas éticas que dirijan el desarrollo tecnológico hacia el bien social.

Este libro forma parte de un proyecto de investigación más amplio sobre las transformaciones de la agencia que devinieron con la revolución digital. Se trata de la primera parte del cuarto volumen de una tetralogía que el autor denomina *Principia Philosophiae Informationis*, que incluye *The Philosophy of Information* (2011), *The Ethics of Information* (2013) y *The Logic of Information* (2019). La segunda parte que continuará este cuarto volumen será su próximo libro *La política de la información*. En palabras del propio autor, "la tarea de este volumen sigue siendo contribuir, como en los anteriores, al desarrollo de una filosofía de nuestro tiempo para nuestro tiempo" (p. 23).

Este trabajo está dividido en dos partes en dirección a dos objetivos. En la primera parte se persigue un objetivo metateórico, ofreciendo una interpretación filosófica de la IA como tecnología, en donde "la tesis central que allí se desarrolla es que la IA es un divorcio sin precedentes entre agencia e inteligencia" (p. 24). En la segunda parte, el objetivo es la investigación teórica de las consecuencias de aquel divorcio, sistematizando principios y recomendaciones éticas en vista al desarrollo y aplicación de las tecnologías inteligentes de manera coherente y consistente con el desarrollo del bien común.

Primera parte: entendiendo la IA

La primera parte está dividida en tres capítulos, (I) pasado, (II) presente y (III) futuro de la IA, desde el surgimiento de la IA, pasando por su caracterización como nueva forma de agencia, hasta su desarrollo previsible. La hipótesis de esta primera parte, y en la que se basa el desarrollo general del libro, es que "la IA no es una nueva forma de inteligencia, sino una nueva forma de agencia" (p. 137). Los factores principales para este planteamiento son la disociación entre agencia e inteligencia y la *envoltura* de esta agencia disociada a través de entornos cada vez más favorables a la IA.

Floridi parte de la base de que lo digital, en cuanto tecnología de tercer orden —es decir, no solo permite acceder a la información (primer orden, como la imprenta), o procesarla (segundo orden, como el computador), sino que también transforma la manera en que interactuamos con ella y la entendemos, esto es, tiene un importante «poder de corte». Lo digital "transforma radicalmente la realidad porque crea nuevos entornos que habitamos y nuevas formas de agencia con las que interactuamos" (p. 57). Lo digital «acopla, desacopla o reacopla» características del mundo (nuestra ontología) y, por tanto, nuestras asunciones sobre el mundo (nuestra epistemología); lo digital, de esta manera, está reontologizando y reepistemologizando nuestra realidad (pp. 51-59). Para Floridi, hay acoplamiento, por ejemplo, entre la identidad propia y los datos personales, indistinguibles cuando hablamos de la «identidad personal» de los «sujetos de los datos». La «localización» y la «presencia», por otra parte, sufrirían un desacoplamiento, en virtud de la forma en cómo *habitamos* el mundo digital, paralelamente a nuestra disposición física.

En este contexto, para el pensador italiano, existe una "reontologización de la agencia que aún no ha sido seguida de una reepistemologización adecuada de su interpretación" (p. 58). Sin embargo, no estamos hablando aquí de una agencia que implica estados mentales, sino de una agencia informática que procesa datos, actúa conforme a ello de manera autónoma y aprende de sus interacciones. Lo re-

volucionario es que este nuevo tipo de agencia ha "«desacoplado» la capacidad de resolver un problema o completar una tarea con éxito de cualquier necesidad de ser inteligente para hacerlo" (p. 62).

Para comprender mejor esto, es importante advertir que la definición de IA que utiliza Floridi es la que proponen McCarthy, Minsky, Rochester y Shannon en su «Propuesta para un proyecto de investigación de verano sobre inteligencia artificial en Dartmouth» (1955), a saber: "para el presente propósito, el problema de la inteligencia artificial es el de hacer que una máquina se comporte de un modo que se llamaría inteligente si un ser humano se comportara de ese modo" (como se citó en Floridi, 2024, p. 63). Es a través de este enfoque como cobra sentido la tesis propuesta, el triunfo de este nuevo tipo de agencia es el desvincular («desacoplar») la resolución de problemas y la realización de tareas con éxito, de la inteligencia —en un sentido cognitivo, IA Fuerte, en términos de Searle (1980)—.

Como podemos notar aquí, la definición de IA del proyecto de 1955 en Dartmouth se basa en una idea pragmática, y un tanto ambigua, de inteligencia. En efecto, como señala Floridi, "a diferencia de «triángulo», «planeta», «mamífero», la noción de IA probablemente no sea un concepto científico" (p. 70). El tratamiento que se le da al concepto de inteligencia desde la ingeniería suele ser más casuista, es decir, a pesar de no poder definirla en términos estrictos, se tiende a asumir que uno la reconocería si la viera (Floridi, 2024). Esta definición contrafáctica de inteligencia está en línea con las ideas de Turing y el juego de la imitación, que, más allá de responder a la pregunta "¿pueden las máquinas pensar?", apunta a la consideración inteligente de un comportamiento específico. Esta caracterización contrafactual de la IA "contiene las semillas de un enfoque ingenieril (por oposición a uno cognitivo) de la IA" (p. 77).

Para Floridi, estas son las dos almas de la IA, por un lado, la ingenieril, que apunta a «reproducir» inteligencia (tecnologías inteligentes) y, por otro, la cognitiva, cuyo objetivo es «producir» (verdadera) inteligencia. De modo que,

> En lo que respecta al ámbito de la IA [como nueva forma de agencia], lo que importa es el resultado, no si el agente o su comportamiento son inteligentes. Por lo tanto, la IA no consiste en reproducir ningún tipo de inteligencia biológica. Se trata, por otra parte, de intentar prescindir de ella. Las máquinas actuales tienen la inteligencia de una tostadora, y realmente no tenemos mucha idea de cómo pasar de ahí (pp. 82-83)

En este contexto, en opinión del autor, el éxito de la IA se debe principalmente al hecho de que estamos construyendo, a través de los datos, un entorno favorable a la IA, un escenario en donde las tecnologías inteligentes se encuentran "como en casa". Es el mundo el que se está adaptando a la IA y no al revés. Esta es la idea detrás del término *envoltura*: "hoy en día, envolver el entorno en una infoesfera favorable a la IA ha empezado a impregnar todos los aspectos de la realidad. (...) Transformando los hábitats sociales en lugares donde los robots pueden operar con éxito" (p. 88). Gracias al «desacoplamiento» y a la «envoltura», sostiene Floridi, las tecnologías digitales actuales pueden conquistar sin descanso el espacio ilimitado de los problemas y las tareas, siempre que puedan realizarse sin la necesidad de algún "ingrediente" que contribuya a crear lo que llamamos inteligencia humana.

Frente a este panorama, el futuro previsible de la IA depende en buena medida de los datos y, más específicamente, de los tipos de datos con los que trabaja y que produce. Para entender cómo una IA produce sus propios datos es necesario distinguir entre las reglas restrictivas —que señalan las actuaciones ilegales en un marco determinado (como las reglas del fútbol) pero que no determinan *a priori* todos los movimientos posibles— y las reglas constitutivas —que determinan a cabalidad las actuaciones válidas dentro del proceso que regulan (como las reglas del ajedrez)—, distinguiendo, así, datos *históricos*, *sintetizados*, *híbridos* y *sintéticos* (p. 108).

Los datos sintéticos producidos por una IA son aquellos en donde la respuesta o curso de acción se determina de manera autónoma,

no programada, por un algoritmo a partir de las reglas restrictivas observadas inductivamente en un sistema, o a partir de las reglas constitutivamente determinadas que enmarcan el plano de acción. A partir de aquí el concepto clave es el de *ludificación*:

> Solo cuando un proceso o interacción pueda transformarse en un juego, y el juego pueda transformarse en un juego de *reglas constitutivas*, entonces la IA podrá generar sus propios datos, totalmente sintéticos, y ser el mejor «jugador» del planeta. (...) La tendencia hacia la generación de datos tan sintéticos como sea posible (desde sintetizados, más o menos híbridos, hasta totalmente sintéticos) probablemente sea uno de los santos griales de la IA. (pp. 109-111)

Para comprender el desarrollo exponencial de esta agencia inteligente, Floridi destaca además la diferenciación entre problemas difíciles, problemas complejos y la consecuente necesidad de esta envoltura. En efecto, el futuro de la IA no solo depende de la sintetización de los datos, sino que también de la transformación (envoltura) de tareas *difíciles* en tareas *complejas*. Por muy inteligentes que parezcan los artefactos actuales, no son los mejores a la hora de realizar tareas o resolver problemas que requieran un alto grado de destreza como, por ejemplo, la que necesitaría un androide autónomo para conducir un auto común, manipular el volante, los pedales, mirar por los espejos, etcétera. Es por esto que, en opinión del autor, "no estamos construyendo vehículos autónomos poniendo androides en el asiento del conductor, sino repensando todo el ecosistema de vehículos con sus respectivos entornos, (...) eliminando por completo el asiento del conductor" (p. 119). Esto es un buen ejemplo de cómo, al envolver el entorno, estamos transformando un escenario difícil (enemigo de la IA), en uno complejo (amigo de la IA), en donde los recursos computacionales (memoria y pasos necesarios) pueden desplegarse en plenitud.

Finalmente, para cerrar esta primera parte de libro, Floridi nos reitera su tesis central: "Hemos liberado la agencia de la inteligencia" (p. 128). El autor es enfático al señalar que, por ejemplo, los "grandes modelos lingüísticos", *large language models* (Chat GPT), "no razonan

ni comprenden, no son un paso hacia ninguna IA de ciencia ficción, y no tienen nada que ver con los procesos cognitivos presentes en el mundo animal y, sobre todo, en el cerebro y la mente humana, en lo referente a gestionar contenidos semánticos con éxito" (p. 121). De lo que se trata en definitiva el futuro de la IA es de diseño, "ludificar y envolver es cuestión de *diseñar*, o a veces *rediseñar*, las realidades con las que tratamos" (pp. 130).

Segunda parte: evaluando la IA

Esta segunda parte del libro aborda temas específicos de ética de la inteligencia artificial. Luego de un análisis comparativo de las distintas regulaciones existentes en Europa y Norteamérica en materia de IA, presenta un marco unificado de cinco principios éticos para la IA: (1) Beneficencia: promover el bienestar, preservar la dignidad y mantener el planeta; (2) No maleficencia: privacidad, seguridad y «precaución con las capacidades»; (3) Autonomía: el poder de «decidir decidir», es decir, conservar el liderazgo humano frente a la agencia cada vez más autónoma de la tecnologías inteligentes; (4) Justicia: promover la prosperidad, preservar la solidaridad y evitar la injusticia; y (5) Explicabilidad: esto permite los otros principios mediante la inteligibilidad de las tecnologías inteligentes, así como la posibilidad de identificar las responsabilidades por sus actos. A diferencia de los otros principios, que provienen de la bioética, este último es específico de la IA en cuanto tecnología de alta complejidad.

Como es de esperar, el uso de las nuevas tecnologías inteligentes conlleva varios riesgos y problemas. Según Floridi, la mayor parte de estos riesgos surgen precisamente por hacer caso omiso a los principios éticos, mientras que otros problemas estarían directamente relacionados con la estructura misma en que funcionan estas tecnologías, es decir, con los algoritmos. Entre los riesgos de no ser éticos, el autor destaca, por ejemplo, la compra de éticas digitales —elegir o adaptar principios o reglas éticas para justificar prácticas *a posterio-*

ri—, el blanqueamiento de la ética —lavado «azul» (en referencia al *greenwashing* de las compañías para parecer más ecológicos de lo que en realidad son) con afirmaciones infundadas o medidas superficiales—, o el *dumping* ético —exportar actividades de investigación sobre procesos digitales que serían no éticas de realizar localmente e importar sus resultados—. Para Floridi, lo importante es potenciar un enfoque preventivo, en donde "la solución suele ser más y mejor información para todos" (p. 176).

En línea con este enfoque preventivo, para el autor, lo que más importa es cómo se diseña la *infoesfera* y las sociedades de la información maduras que se desarrollan en ella. Los conceptos importantes aquí son ética, *regulación* y *gobernanza*. Sin embargo, el cumplimiento llano de una regulación es ciertamente necesario pero insuficiente, por esto el autor propone la idea de una ética blanda —en oposición a la ética dura que establece los principios generales en los que se basa una regulación—, orientada precisamente a la adaptación específica posregulación, en orden a ir más allá de lo que puede exigir la ley en un caso concreto.

Por otra parte, los algoritmos también representan peligros importantes al no ser éticamente neutrales. Estos problemas éticos se basan en que los algoritmos pueden utilizarse para convertir datos en pruebas para un determinado resultado, y de ese modo desencadenar y motivar una acción que puede tener consecuencias éticas. Esto se relaciona, a su vez, con la trazabilidad, es decir, con la posibilidad de atribuir las responsabilidades por los efectos de las acciones que puede desencadenar el uso de un algoritmo (pp. 205-208). La evidencia no concluyente, inescrutable o equivocada es parte de los riesgos que conlleva el uso de algoritmos para tomar decisiones o asignar recursos. Problemas específicos como la *apofenia* —el ver patrones donde en verdad no los hay—, recuerdan que la correlación no implica necesariamente causalidad, los patrones pueden ser el resultado de propiedades inherentes del sistema modelado por los datos, de los conjuntos de datos, o de la hábil manipulación de los conjuntos de datos (p. 210).

A partir de aquí, el autor presenta una revisión de las buenas y malas prácticas en el uso de la IA, sus riesgos y oportunidades, así como también una serie de recomendaciones para una buena sociedad de la IA. Entre las malas prácticas destaca el uso delictivo de la IA, además de la desviación de responsabilidades causales y morales por el uso malicioso o daño colateral que se pueda provocar con ella. Por otra parte, la idea de utilizar las tecnologías inteligentes para el bien común supone contribuir a reducir, mitigar o erradicar un determinado problema social o medioambiental sin producir nuevos daños o amplificar los existentes (p. 300). Es muy importante para el éxito en el uso de la IA evitar que los sesgos que puedan estar presentes en los datos durante su desarrollo se traspasen a la toma de decisiones algorítmicas y, en específico, que esos prejuicios se arraiguen, refuercen y perpetúen de nuevo mediante mecanismos erróneos de aprendizaje reforzado (p. 331).

Floridi apuesta por diseñar y construir una buena sociedad de la IA, aprovechando sus virtudes para combatir problemas tan graves como el cambio climático, permitiendo la autorrealización y agencia humana, aumentar las capacidades de la sociedad y cultivar la cohesión social. Sin desestimar, por cierto, los riesgos de la infrautilización de estas nuevas tecnologías, así como tampoco su sobreuso o mala utilización, lo que podría erosionar en definitiva la autodeterminación humana. La infoesfera es este nuevo entorno informativo que las tecnologías de los datos y la comunicación están creando, y el que cada día habitamos con mayor frecuencia. En este escenario, es interesante advertir que el éxito sin precedentes de las nuevas tecnologías digitales se debe a que son ellas las verdaderas nativas de la infoesfera: nosotros "somos meros organismos analógicos que intentan adaptarse a un hábitat tan nuevo simplemente viviendo *onlife*" (p. 408).

Comentarios finales

El abundante trabajo de Floridi en torno a la información y las tecnologías inteligentes, así como la reflexión sobre las consecuencias

éticas de su masificación, son una valiosa base teórica desde la cual analizar nuestro tiempo presente y trazar las directrices del futuro. Y en este mismo sentido, a partir de las propias ideas del filósofo italiano, surgen amplios terrenos fértiles para el debate, la crítica y la reflexión.

En efecto, la tesis central del libro sobre este divorcio entre agencia e inteligencia resulta inquietante, pues deja en suspenso alguna definición clara de esta inteligencia humana soslayada por la tecnología más allá de señalar que hace referencia a una visión cognitiva. El enfoque desde el que Floridi aborda la problemática de la ética de la IA se basa precisamente en aquel «desacoplamiento» entre agencia e inteligencia. Pero cabe preguntar, ¿qué o cuál es, o en qué consiste este "ingrediente" que se deja fuera, que se divorcia de la agencia? Esta pregunta es relevante y tiene importantes consecuencias con respecto a la propuesta ética del autor.

Floridi es tajante al señalar que esta nueva agencia de la tecnología no es *realmente* inteligente y que depende de la envoltura, que "no estamos construyendo vehículos autónomos poniendo androides en el asiento del conductor" (p. 119). Esto puede ser cierto cuando hablamos de automóviles, pero la verdad es que sí existen modelos robóticos y androides capaces de un desplazamiento autónomo y de motricidad similar a la humana, como los presentados en los últimos años por Boston Dynamics o el modelo Optimus de Tesla. El concepto de envoltura es brillante, pero no es el único camino por el que avanza la tecnología.

Además, el concepto de IA que Floridi toma de la propuesta de Dartmouth de 1955 tiene problemas profundos que quedan en el aire, al estar basado en el enfoque ingenieril y pragmático inspirado en el "juego de la imitación" de Alan Turing (1950), es una definición que precisamente se establece con el objetivo de evitar la discusión de qué es la inteligencia o qué es una máquina. Sin embargo, esta definición contrafáctica de la IA parece depender de la consideración específica de lo que entendamos por «comportamiento que si realizara un hu-

mano sería considerado comportamiento inteligente». ¿Es posible un comportamiento específicamente humano *no inteligente*? Al entender la IA según el comportamiento que en un humano se consideraría inteligente, cabe la visión contrapuesta: ¿qué comportamiento de un humano sería considerado comportamiento mecánico? Precisamente el no humano, precisamente el comportamiento que compartimos con otros seres vivos, aquel que descartamos a la hora de definirnos como seres humanos y que no sirve, en definitiva, para decidir la inteligencia de una máquina.

A simple vista, no puede haber un comportamiento específicamente humano que a la vez no sea inteligente. El definir la IA en virtud de un comportamiento que si realizara una persona sería considerado inteligente conlleva el peligro de una petición de principio, ya que, siendo el ser humano inteligente por definición, se haga lo que se haga, si es hecho por un ser humano, lo podemos considerar comportamiento inteligente precisamente porque de antemano ya hemos aceptado esta característica como intrínsecamente humana. Parece ser que el "ingrediente" de la inteligencia cognitiva, de la "verdadera" inteligencia, es un fenómeno que no es directamente observable (Maturana, 2013, p. 24) y, sin embargo, Floridi sostiene que se ha «desacoplado» de la agencia.

Por otra parte, si la IA, en cuanto agencia artificial sobre la que trabaja Floridi, no tiene posibilidades de transformarse en Inteligencia Artificial Fuerte, en inteligencia artificial de "ciencia ficción", la autodeterminación humana no estaría más amenazada de lo que ya pueda estar y tampoco deberían haber tantos problemas éticos nuevos como el autor nos sugiere, puesto que estaría claro que la responsabilidad por su mal uso debiera estar en todo el espectro de desarrolladores y aplicadores de estas tecnologías, así como en cualquier otra. La distribución de responsabilidades no debería ser distinta a la que tiene lugar hoy en día ante la producción de artículos peligrosos o consecuencias nocivas derivadas de su producción para cualquier comunidad. El hecho de que no se puedan proyectar todas, absolutamente todas las consecuencias de las aplicaciones de una tecnología, no es extraño, de

hecho, así funciona la aplicación de justicia en general, *a posteriori*. La ley en el Estado moderno de derecho establece un marco ideal en el despliegue social, sin embargo, es imposible que prevea todas las formas en que los objetivos deseables pueden ser amenazados, de modo que son los tribunales los llamados a interpretar y aplicar la ley de manera coherente. De otro modo, se presenta el riesgo de una hiperregulación normativa con todo lo que eso conlleva.

El sesgo en los datos, en cuanto riesgo del uso de la IA, no es algo nuevo en ciencias sociales como la economía o la sociología. La falta de transparencia en los procesos industriales en cuanto principio ético tampoco es algo que solo afecte a las agencias artificiales, y los demás principios ya son parte de la mayoría de los sistemas jurídicos occidentales. Por ejemplo, que determinadas industrias no hubieran previsto su aporte negativo al calentamiento global no las hace menos responsables. Podemos pensar lo mismo de la industria de la tecnología de tercer orden.

Por el momento la IA Fuerte es ciencia ficción y sospecho que seguirá siéndolo por un buen tiempo. En la medida en que ni siquiera es posible establecer con precisión los límites y características de la inteligencia como tal, es improbable que podamos estar satisfechos con los intentos de producirla. No obstante, parece ser esta vertiente cognitiva la que más excitación y preocupación provoca, y al descartarla de plano y tratar a la IA como mero sistema de resultados, de consecución de tareas y solución de problemas, Floridi amenaza peligrosamente en sus fundamentos la necesidad de una ética radicalmente vanguardista para la regulación de una herramienta tecnológica que cualitativamente no sería mucho mejor que "una tostadora", lo que precisamente parece ser contrario a todo lo que nos propone después.

Referencias bibliográficas

Floridi, L. (2011). *The Philosophy of Information*. Oxford University Press.

Floridi, L. (2013). *The Ethics of Information*. Oxford University Press.

Floridi, L. (2019). *The Logic of Information: A Theory of Philosophy as Conceptual Design*. Oxford University Press.

Floridi, L. (2024). Ética de la inteligencia artificial. Herder.

González, R. (2016). El entendimiento lingüístico en la Inteligencia Artificial: una relación ambivalente con Descartes. *Revista IF Sophia*, *2*(7), 1-32.

Hui, Y. (2022). *Recursividad y contingencia*. Caja Negra.

Maturana, H. (2013). *Desde la biología a la psicología*. Editorial Universitaria.

Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences, 3*(3), 417-424.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind* (49), 433-460.