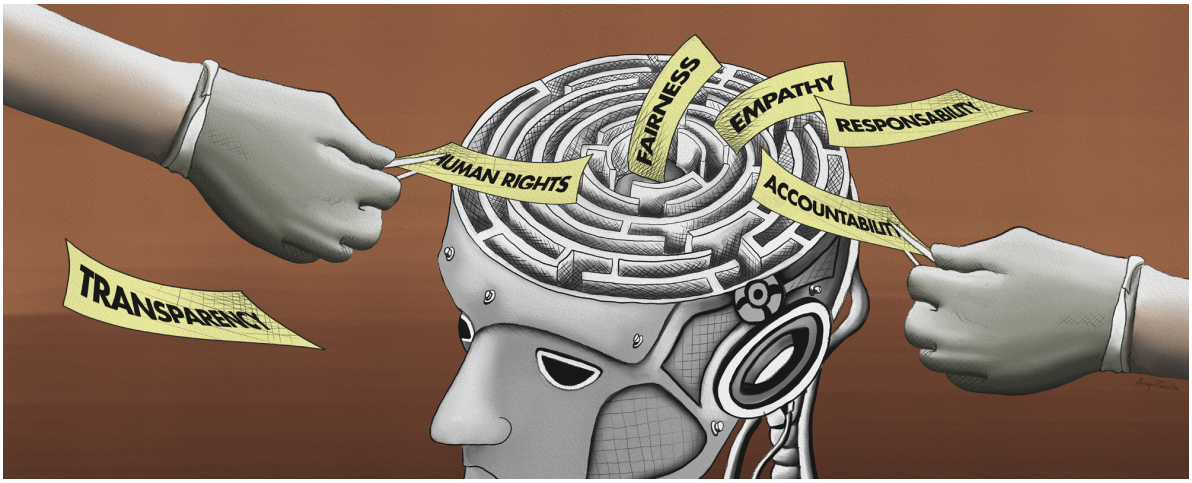
RETOS ÉTICOS

Reflexión sobre la «ética» en la ética de la IA

Pak-Hang Wong, Judith Simon



Araya Peralta

A principios de 2019 una gran consultora internacional identificó al «ético de la IA» como un cargo fundamental para que las empresas apliquen correctamente la inteligencia artificial (IA). Sostiene que los éticos de la IA son necesarios para ayudar a las empresas a explorar las cuestiones éticas y sociales que plantea el uso de la IA [1]. La opinión de que la IA es positiva y, aun así, potencialmente *peligrosa* para los individuos y la sociedad es muy compartida por el sector, los círculos académicos, los gobiernos y las organizaciones de la sociedad civil. En consecuencia, y con el fin de materializar las ventajas y al mismo tiempo esquivar los obstáculos éticos y los efectos nefastos, se han puesto en marcha numerosas iniciativas para) examinar las dimensiones ética, social, jurídica y política de la IA y b) elaborar unas directrices y recomendaciones éticas para el diseño y la aplicación de la IA [2]

.

Sin embargo, a veces las cuestiones terminológicas dificultan el análisis profundo de los problemas éticos de la IA. Las definiciones de *inteligencia* e *inteligencia artificial* a menudo son imprecisas y las diferentes interpretaciones de estos términos destacan diferentes inquietudes. Para evitar la confusión y el riesgo de que todo acabe siendo un diálogo de sordos, cualquier debate constructivo sobre la ética de la IA exige que se explique la definición d'IA que se utiliza y también que se especifique el tipo d'IA sobre el cual se debate. Con respecto a la definición, consultamos el Grupo Europeo de Ética de la Ciencia y de las Nuevas Tecnologías de la Comisión Europea, que define la IA como los «sistemas de software (y posiblemente también de hardware) diseñados por seres humanos que, dado un

objetivo complejo, actúan en la dimensión física o digital percibiendo su entorno mediante la adquisición de datos, interpretando los [datos…]recopilados, razonando sobre el conocimiento o procesando la información derivada de estos datos y decidiendo cuáles son las mejores medidas que hay que tomar para alcanzar un objetivo determinado. Los sistemas d'IA pueden utilizar reglas simbólicas o aprender un modelo numérico y también pueden adaptar su conducta analizando el impacto de sus acciones anteriores sobre el entorno» [3].

Para ofrecer orientación y recomendaciones específicas, el análisis ético de la IA también tiene que especificar la *tecnología* (por ejemplo, vehículos autónomos, sistemas de recomendación, etc.), los *métodos* (por ejemplo, aprendizaje profundo, aprendizaje de refuerzo, etc.) y los *sectores de aplicación* (por ejemplo, atención sanitaria, finanzas, noticias, etc.). En este artículo nos centraremos en las cuestiones éticas relacionadas con la *IA autónoma*, es decir, los agentes artificiales que pueden decidir y actuar independientemente de la intervención humana, e ilustraremos las cuestiones éticas de la IA autónoma con un gran número de ejemplos.

Fijémonos primero en el caso de los vehículos autónomos (VA). La posibilidad de que haya escenarios de accidentes con VA implicados, en el que inevitablemente pasajeros o peatones resultarían heridos, ha obligado los investigadores y desarrolladores a reflexionar sobre cuestiones relativas a la aceptabilidad ética de las decisiones tomadas por los VA, como qué decisiones tienen que tomar los VA en estos casos, cómo se pueden justificar estas decisiones, qué valores reflejan los VA y sus decisiones, etc. [4].

> Los algoritmos de contratación acostumbran a funcionar aplicando los criterios que han aprendido de un conjunto de datos de entrenamiento. Lamentablemente, estos datos de entrenamiento pueden ser sesgados y dar lugar a modelos potencialmente discriminatorios.

Todavía mejor, fijémonos en el caso de los algoritmos de contratación, que se han introducido para automatizar el proceso de recomendación, preselección y quizás incluso selección de candidatos de un trabajo. Los algoritmos de contratación acostumbran a funcionar aplicando los criterios que han aprendido de un conjunto de datos de entrenamiento. Lamentablemente, estos datos de entrenamiento pueden ser sesgados y dar lugar a modelos potencialmente discriminatorios [5].

Con el fin de garantizar la protección contra la discriminación, que no sólo es un derecho humano sino que está incluida en las constituciones de muchos países, nos tenemos que asegurar de que estos algoritmos son, como mínimo, no discriminatorios e idealmente también *equitativos*. Sin embargo, la equidad tiene diferentes interpretaciones: las personas discrepan sobre el significado de la equidad, pero además el concepto de equidad

también puede depender del contexto. Todavía más, se ha demostrado que no se pueden obtener diferentes parámetros de equidad de manera simultánea [6]. Eso plantea la cuestión de cómo se tienen que concebir valores como la equidad en qué contexto y cómo se pueden aplicar.

Por lo tanto, una de las cuestiones fundamentales de la ética de la IA se puede formular como un problema de alineación de valores: cómo podemos crear una IA autónoma alineada con los valores arraigados en la sociedad [7]. Virginia Dignum ha distinguido tres dimensiones de la ética de la IA: el «ética por diseño», el «ética en el diseño» y el «ética para el diseño» [8], que sirven para identificar dos respuestas diferentes al problema de la alineación de valores. Estructuraremos el análisis siguiente de acuerdo con las tres dimensiones mencionadas y exploraremos más detalladamente las dos vías diferentes para responder al problema de la alineación de valores.

Crear una IA ética: perspectivas y limitaciones

La ética por diseño es «la integración técnica/algorítmica de la capacidad de razonamiento como parte de la conducta de [la IA autónoma]» [9]. Esta línea de investigación también se llama «ética de las máquinas» y aspira a crear agentes morales artificiales, que son agentes artificiales con capacidad ética y pueden tomar decisiones éticas sin la intervención humana [10]. Así pues, la ética de las máquinas responde al problema de la alineación de valores con la creación de una IA autónoma que por sí sola se alinea con los valores humanos. Ilustramos esta perspectiva con los ejemplos de los VA y los algoritmos de contratación: los investigadores y desarrolladores se esfuerzan por crear VA que puedan razonar sobre la decisión moralmente correcta y actuar en consecuencia en escenarios en que se provoque un daño inevitable. Asimismo, se supone que los algoritmos de contratación toman decisiones no discriminatorias sin la intervención humana.

Wendell Wallach y Colin Allen definieron tres tipos de aproximaciones a la ética de las máquinas en su libro seminal *Moral machines* [11]. Los tres tipos de aproximaciones son, respectivamente (y) aproximaciones descendentes, (ii) aproximaciones ascendentes y (iii) aproximaciones híbridas, que combinan las aproximaciones descendentes y ascendentes. En su forma más sencilla, la aproximación descendente intenta formalizar y aplicar una teoría ética concreta a la IA autónoma, mientras que la aproximación ascendente pretende crear una IA autónoma que pueda aprender del entorno o a partir de una serie de ejemplos qué es moralmente correcto o no. En último lugar, la aproximación híbrida combina técnicas y estrategias tanto de la aproximación descendente como de la ascendiente [12].

Estas aproximaciones, sin embargo, están sujetas a algunas limitaciones *teóricas* y *técnicas*. Por ejemplo, las aproximaciones descendentes tienen que superar el reto de encontrar y defender una teoría ética incontrovertible entre tradiciones filosóficas *contradictorias*. Si no, la IA ética corre el riesgo de construirse sobre una base *inadecuada* o incluso *falsa*. Por otra parte, las aproximaciones ascendentes deducen qué es ético a partir de aquello que es *popular* o que se considera *universalmente* ético en el contexto o entre ejemplos. Con todo, estas deducciones no garantizan que la IA autónoma adquiera reglas o principios éticos

genuinos, porque ni la popularidad ni el hecho de que algo se considere ético ofrecen una *justificación* ética adecuada [13]. Además, está la dificultad técnica en crear una IA ética que pueda discernir con eficacia la información éticamente relevante de la éticamente irrelevante entre la gran cantidad de información disponible en un contexto determinado. Esta capacidad sería necesaria tanto para la aplicación correcta de los principios éticos en las aproximaciones descendentes como para la adquisición correcta de los principios éticos en las aproximaciones ascendentes [14].

> La IA autónoma en general y la IA ética en particular pueden erosionar considerablemente la autonomía humana porque las decisiones que toman por nosotros o sobre nosotros estarán fuera de nuestro control.

Aparte de los retos teóricos y técnicos, la creación de IA autónoma con capacidades éticas ha recibido a algunas críticas éticas. En primer lugar, la IA autónoma en general y la IA ética en particular pueden erosionar considerablemente la autonomía humana porque las decisiones que toman por nosotros o sobre nosotros estarán fuera de nuestro control y, por lo tanto, reducirán nuestra independencia de las influencias externas [15]. En segundo lugar, no queda claro quién o qué tiene que ser responsable de las decisiones erróneas que la IA autónoma tome, lo cual suscita inquietud en relación con sus efectos sobre nuestras prácticas de responsabilidad moral [16]. En último lugar, los investigadores han argumentado que convertir la IA autónoma en agentes morales o pacientes morales complica innecesariamente nuestro mundo moral con la introducción de elementos extraños que son ajenos a nuestra conciencia moral y, en consecuencia, con la imposición de una carga ética innecesaria sobre los seres humanos, ya que nos exige que prestemos una atención moral excesiva a la IA autónoma [17].

La ética de las máquinas, la ética truncada

Nuestro análisis de las dificultades teóricas, técnicas y éticas de la ética de las máquinas no pretende ser exhaustiva ni conclusiva y, en realidad, estas dificultades se podrían superar en la investigación y el desarrollo futuros de la IA autónoma. No obstante, creemos que estas dificultades se merecen que nos detengamos y reconsideremos las posibilidades de crear una IA ética. De hecho, queremos proponer una crítica más fundamental de la ética de las máquinas antes de explorar otra vía para responder al problema de alineación de valores.

Recordemos que el objetivo de la ética de las máquinas es crear una IA autónoma que pueda tomar decisiones éticas y actuar con ética sin la intervención humana. Se centra en infundir a la IA autónoma la capacidad de tomar decisiones éticas y hacer acciones éticas, cosa que refleja una interpretación peculiar de la «ética» que escogemos cuestionar. Más concretamente, si se centra *sólo* en la capacidad de decidir y actuar con ética, la ética de

las máquinas se expone a una visión *truncada* de la ética que considera que las decisiones y acciones éticas se pueden separar de sus contextos sociales y relacionales. La filósofa y novelista Iris Murdoch, por ejemplo, hace mucho tiempo argumentaba que la moralidad no tiene nada que ver con «una serie de elecciones explícitas que se producen en una serie de situaciones especificables» [18], sino con la «autorreflexión o actitudes complejas ante la vida que se manifiestan y se elaboran continuamente en el discurso público y privado, pero no se pueden separar temporalmente en situaciones» [19]. Para Murdoch, aquello que es ético está inherentemente vinculado a un trasfondo de valores. Por eso, a la hora de reflexionar sobre la ética, es fundamental mirar *más allá* de la capacidad de decidir y actuar con ética y los momentos de la decisión y la acción éticas, y fijarse en el trasfondo de valores y los relatos que hay detrás de la decisión y la acción. Se han esgrimido argumentos parecidos para reafirmar la función de los contextos sociales y relacionales a la hora de limitar las decisiones éticas y configurar las consecuencias morales y, por lo tanto, la importancia de tenerlos en cuenta en nuestra reflexión ética [20].

En esta misma línea crítica, el énfasis para infundir las capacidades éticas de la IA a la ética de las máquinas se puede considerar una equivocación si este énfasis eclipsa el hecho de que las consecuencias éticas de la IA autónoma están determinadas por varios factores interconectados ajenos a su capacidad de razonamiento ético, y que hay un proceso dilatado de negociación social y política sobre los criterios de corrección e incorrección que señalan las posibles decisiones y acciones éticas de la IA autónoma. El experimento de la Máquina Moral, llevado a cabo por investigadores del MIT Media Lab, es un buen ejemplo [21]. En el experimento, los investigadores del MIT intentan externalizar las decisiones éticas en diferentes escenarios de accidentes con VA implicados y los resultados tienen que servir para orientar el diseño ético de los VA. Sin embargo, faltan el trasfondo social, cultural y político y las historias personales de accidentes *reales* que los escenarios de accidentes del experimento no describen —y a menudo no pueden describir— adecuadamente [22]. En este sentido, el experimento de la Máquina Moral también se basa en una visión truncada de la ética, en que *sólo* se tiene en cuenta la decisión de que se tiene que tomar en situaciones concretas y olvida el trasfondo de valores y los datos contextuales cruciales para tomar decisiones éticas.

> A la hora de reflexionar sobre la ética, es fundamental mirar más allá de la capacidad de decidir y actuar con ética y los momentos de la decisión y la acción éticas, y fijarse en el trasfondo de valores y los relatos que hay detrás de la decisión y la acción

De hecho, los contextos sociales y relacionales son importantes para el análisis ético de la IA autónoma *antes* y *después* de su aplicación. Por ejemplo, se puede diseñar un algoritmo de contratación imparcial, que valore a los candidatos a un puesto de trabajo basándose *sólo* en las cualidades requeridas en una oferta. Aun así, este algoritmo de contratación imparcial podría seguir siendo discriminatorio, y por lo tanto éticamente discutible, si las cualidades específicas requeridas en una oferta están ligadas involuntariamente a la raza,

el género y la clase social. En este caso, hay que ir con cuidado para no reproducir el *sesgo social preexistente* en el algoritmo de contratación. Además, incluso las tecnologías mejor intencionadas pueden afectar muy negativamente los (no) usuarios, ya que el sesgo y el perjuicio podrían *surgir* de la interacción entre la tecnología y los usuarios y la sociedad [23]. Imaginad una aplicación que los ciudadanos pueden utilizar para denunciar incidentes, como infraestructuras deterioradas, en el ayuntamiento de la ciudad, que a su vez utiliza un algoritmo para ordenar y clasificar los problemas locales según estas denuncias. Si damos por hecho que el acceso a los teléfonos inteligentes y, por lo tanto, a la aplicación tiene una distribución desigual, eso puede significar una falta de denuncias en las zonas con ciudadanos más pobres. Si este sesgo de los datos de entrada no se tiene en cuenta en la ordenación y la clasificación algorítmicas, las desigualdades entre las zonas más ricas y más pobres de la ciudad podrían crecer todavía más [24].

La lección principal de estos dos ejemplos es que incluir unas cuantas normas y principios éticos en la IA autónoma es insuficiente para resolver el problema de alineación de valores, porque los trasfondos y los contextos *sí* que contribuyen a nuestra posición global sobre la cual es ético. Nos tenemos que acordar de que la IA autónoma *siempre* se sitúa en unos contextos sociales y relacionales más amplios y que, por eso, no nos podemos centrar *sólo* en su *capacidad* en tomar decisiones y actuar con ética. Tenemos que tener en cuenta no sólo *qué* decisiones y acciones tiene que hacer la IA autónoma, sino también (y) *por qué* nosotros (o, mejor dicho, la sociedad) creemos que estas decisiones y acciones son éticas, (ii) *cómo* llegamos a estas conclusiones y (iii) *si* tenemos justificación para pensar así. Así pues, el experimento de la Máquina Moral es cuestionable porque supone injustificablemente que la respuesta más *intuitiva* o *popular* a los escenarios de accidentes es la respuesta *ética*. En realidad, la reformulación de las preguntas nos ofrece dos ventajas. Primero, ahora podemos incluir fácilmente otras partes y factores *además de* la IA autónoma en nuestra reflexión ética. Segon, también explicita la posibilidad de (re)negociar qué reglas o qué principios éticos se tienen que incluir en la IA autónoma (o incluso cuestionar el uso de la IA autónoma en un contexto concreto).

Una ética distribuida de la IA

Seamos claros: no neguemos la necesidad de examinar los valores incorporados a la tecnología ni la importancia de diseñar y crear tecnología con unos valores alineados con los intereses humanos [25]. Tal como demuestran los ejemplos de este artículo, la IA autónoma puede tener un papel en la toma de decisiones éticas y ocasionar unos efectos éticamente relevantes. Por eso, hace falta examinar los valores que incorpora y utilizar los valores sociales compartidos para orientar el diseño y el desarrollo. No obstante, queremos cuestionar el deseo de *delegar* el razonamiento ético y la decisión a las máquinas y, así, arrancar el razonamiento y la decisión de los contextos sociales y relacionales. Un buen análisis de la ética de la IA tiene que ampliar su ámbito de reflexión e incluir otras partes y factores que sean relevantes para la toma de decisiones éticas y hayan contribuido a los efectos éticos de la IA autónoma. Para eso, es fundamental que la ética de la IA incluya diferentes partes interesadas (como responsables políticos, directivos de empresas, diseñadores, ingenieros, usuarios, no usuarios y el público en general) en la reflexión ética

sobre la IA autónoma. De hecho, sólo si lo hacemos así podremos dar una respuesta adecuada a las preguntas siguientes: (i) *por qué* creemos que las decisiones y los efectos de la IA son éticos, (ii) *cómo* llegamos a estas conclusiones y (iii) *si* tenemos justificación para opinar así.

> En el diseño y la aplicación de la IA deben de tenerse en cuenta las desigualdades y las injusticias sociales existentes, considerarlas y, en el mejor de los casos, intentar atenuarlas mediante las decisiones de diseño

Esta ética ampliada de la IA la denominaremos la *ética distribuida de la IA*. El adjetivo «distribuida» pretende reflejar el hecho de que hay múltiples partes y factores que son relevantes y han contribuido a los efectos éticos de la IA autónoma y que, por lo tanto, la responsabilidad se «distribuye» entre las partes y los factores relevantes y coadyuvantes [26]. Volvemos a los ejemplos de los VA y los algoritmos de contratación: una mala ordenación urbana y unas carreteras deficientes tienen que ser preocupaciones legítimas en la ética de los VA, de la misma manera que los sesgos sociales y culturales existentes son consideraciones válidas para los algoritmos de contratación éticos. Por lo tanto, en el diseño y la aplicación de la IA deben de tenerse en cuenta las desigualdades y las injusticias sociales *existentes*, considerarlas y, en el mejor de los casos, intentar atenuarlas mediante las decisiones de diseño.

La ética distribuida de la IA necesita lo que Dignum denominó «ética *en* el diseño», es decir, «los métodos legislativos y de ingeniería que ayudan a analizar y evaluar las implicaciones éticas de los sistemas d'IA, ya que estos integran o sustituyen las estructuras sociales tradicionales», así como la «ética *para* el diseño», es decir, «los códigos de conducta, las normas y los procesos de certificación que garantizan la integridad de los desarrolladores y los usuarios cuando investigan, diseñan, construyen, utilizan y gestionan sistemas de inteligencia artificial» [27]. Las cuestiones éticas de la IA autónoma no se pueden resolver con una capacidad ética *individual(ista)* «mejor», sino sólo con el *esfuerzo colectivo*. Para orientar este esfuerzo colectivo, las directrices éticas ofrecen recursos prácticos para suscitar una reflexión basada en valores y principios en relación con la IA autónoma y coordinar con eficacia el esfuerzo entre las diferentes partes relevantes y coadyuvantes [28].

Conclusiones: sobre la IA fiable de la UE

En abril del 2019, el grupo de expertos de alto nivel publicó las *Directrices éticas para una IA fiable*, en la que se concreta la visión europea sobre la IA. De acuerdo con estas directrices, Europa tiene que investigar y desarrollar una *IA fiable*, es decir, *lícita*, *ética* y *robusta*.

Hay dos puntos de las directrices que merecen una mención especial en el tema de debate que nos ocupa. En primer lugar, es interesante señalar que la preocupación por la fiabilidad a las directrices concierne «no sólo las propiedades inherentes a esta tecnología, sino también las cualidades de los sistemas sociotécnicos en que se aplica la IA. […] Por lo tanto, los esfuerzos dirigidos a garantizar la fiabilidad de la IA no conciernen sólo la confianza en que suscita el mismo sistema d'IA, sino que requieren un enfoque integral y sistémico que abrace la fiabilidad de todos los agentes y procesos que forman parte del contexto sociotécnico en que se enmarca el sistema a lo largo de todo su ciclo de vida» En este sentido, la idea de una IA fiable coincide claramente con la ética distribuida de la IA que hemos descrito antes. En segundo lugar, también es interesante señalar que los cuatro principios éticos identificados en las directrices son *principios de nivel medio*, es decir:

1. El principio de respeto a la autonomía humana.
2. El principio de prevención de los daños.
3. El principio de equidad.
4. El principio de explicabilidad.

La formulación de principios éticos basados en *principios de nivel medio* es especialmente esclarecedora, ya que los principios de nivel medio *exigen* una interpretación humana y organizar la aplicación y no están pensados para (y de hecho no pueden) aplicarse a la IA autónoma. La necesidad de interpretación y organización también apunta a los contextos sociales y relacionales, donde se encuentran los recursos para la interpretación y la organización.

Aunque la idea europea de una IA fiable y las directrices tienen unos fundamentos sólidos desde la perspectiva conceptual, plantean algunos problemas. Por ejemplo, el uso de principios de nivel medio en las directrices da bastante espacio para la interpretación, cosa que los agentes malintencionados pueden aprovechar para seleccionar las interpretaciones y sacudirse la responsabilidad. Este problema se agrava por el énfasis en el autorregulación que se hace en las directrices. Los políticos y las empresas pueden hacer falsas promesas con medidas *baratas* y *superficiales*, como la propaganda y la creación de consejos asesores simbólicos, sin abordar *sustancialmente* los efectos negativos de la IA. Por lo tanto, hay problemas importantes relacionados con el marco regulador e institucional *real* con respecto a la ética de la IA y la materialización de esta idea europea. En concreto, está la necesidad de crear un marco claro para hacer una distribución *equitativa* de los beneficios y los riesgos de la IA y la necesidad de introducir leyes y normativas «duras» contra la violación de los valores éticos básicos y los derechos humanos.

A pesar de estos problemas, la atención de las directrices *en los humanos* y *más allá de la tecnología* se tiene que considerar una posición *normativa* adecuada para la ética de la IA y la idea europea. Para acabar este artículo, queremos recordar que las cuestiones éticas relativas a la IA autónoma se distribuyen por naturaleza y que nosotros (o, mejor dicho, la sociedad) tenemos que tener voz en su diseño y su aplicación.

REFERÈNCIES

1 — KPMG (2019) Top 5 AI hires companies need to succeed in 2019.
2 — AlgorithmWatch ha elaborado una lista de marcos y directrices éticas disponible en: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/.
3 — European Commission High-Level Expert Group on Artificial Intelligence [AI HLEG] (2019) Ethics guidelines for trustworthy AI. European Commission.
4 — El tipo de escenarios de accidentes se conece como «el dilema de la vagoneta». Es tan solo uno de los temas que se debaten en la ética de los vehículos autónomos y lo ponemos de ejemplo sólo para ilustrar uno de los numerosos problemas éticos que la IA autónoma podria plantear. Véase:
• Lin, P. (2016) Why ethics matters for autonomous cars. In M. Maurer, J. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous Driving: Technical, Legal and Social Aspects* (pp. 69-85). Berlin: Springer.
• Keeling, G. (2019) Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics*.
5 — Bogen, M. (2019) All the ways hiring algorithms can introduce bias. *Harvard Business Review*, May 6, 2019.
6 — Véase:
• Friedler, S., Scheidegger, C., & Venkatasubramanian, S. (2016) On the (Im)possibility of fairness. arXiv:1609.07236.
• Chouldechova, A. (2017). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* 5(2): 153-163.
• Wong, P.-H. (2019) Democratizing algorithmic fairness. *Philosophy & Technology*.
7 — El problema de la alineación de la IA lo formuló explícitamente por primera vez Stuart Russell en 2014. Véase:
• Peterson, M. (2019) The value alignment problem: a geometric approach. *Ethics and Information Technology* 21 (1): 19-28.
8 — Dignum, V. (2018) Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology* 20 (1): 1-3.
Íbid., p. 2
10 — Véase:
• Winfield, A., Michael, K., Pitt, J., & Evers, V. (2019) Machine ethics: the design and governance of ethical AI and autonomous systems. *Proceedings of the IEEE* 107 (3): 509-517.
• Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.
• Misselhorn, C. (2018) Artificial morality. concepts, issues and challenges. *Society* 55 (2): 161-169.
11 — Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.
12 — Íbid., p. 79-81
13 — Puede encontrarse un análisis de la dificultad ética de las máquinas en: Cave, S., Nyrup, R., Vold, K., & Weller, A. (2019) Motivations and risks of machine ethics. *Proceedings of the IEEE* 107 (3): 562-74.

14 — También se conoce como el problema del marco moral. Véase: Horgan, T., & Timmons, M. (2009) What does the frame problem tell us about moral normativity? *Ethical Theory and Moral Practice* 12 (1): 25-51.
15 — Danaher, J. (2018) Toward an ethics of AI assistants: an initial framework. *Philosophy & Technology* 31 (4): 629-653.
16 — Matthias, A. (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6 (3): 175-83.
17 — Bryson, J. J. (2018) Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology* 20 (1): 15-26.
18 — Murdoch, I. (1956) Vision and choice in morality. *Proceedings of the Aristotelian Society, Supplementary* 30: 32-58. p. 34
19 — Íbid., p. 40
20 — Walker, M. U. (2007) *Moral Understandings: A Feminist Study in Ethics*. Oxford: Oxford University Press.
21 — Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018) The Moral Machine experiment. *Nature* 563: 59-64.
Jaques, A. E. (2019) Why the moral machine is a monster. Paper presented to We Robot 2019, University of Miami, April 11-13, 2019.
23 — Friedman, B., & Nissenbaum, H. (1996) Bias in computer systems. *ACM Transactions on Information Systems* 14 (3): 330-347.
24 — Simon J (2012) E-Democracy and Values in Design. *Proceedings of the XXV World Congress of IVR 2012*.
Simon, J. (2017) Value-sensitive design and responsible research and innovation. In S. O. Hansson (Ed.), *The Ethics of Technology Methods and Approaches* (pp. 219-235). London: Rowman & Littlefield.
26 — Véase:
• Floridi, L. (2013) Distributed morality in an information society. *Science and Engineering Ethics* 19 (3): 727-743.
• Simon, J. (2015) Distributed epistemic responsibility in a hyperconnected era. In L. Floridi (Ed.), *The Onlife Manifesto* (pp. 145-159). Cham, Springer.
27 — Dignum, V. (2018) Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology* 20 (1): p. 2.
28 — Floridi, L. (2019) Establishing the rules for building trustworthy. *Nature Machine Intelligence* 1: 261-262.

Pak-Hang Wong
Pak-Hang Wong és graduat en Filosofia i Història per la Universitat de Hong Kong, i doctor per la universitat de Twente. És investigador associat posdoctorat del Grup de Recerca sobre Ètica en Tecnologies de la Informació del Departament d'Informàtica de la Universitat d'Hamburg. El seu àmbit d'interès principal és la filosofia de la tecnologia, on examina els reptes de les tecnologies digitals emergents –Big Data, AI, polítiques d'algoritmes– per comprendre la responsabilitat moral i la pràctica del cultiu de la virtut a través del confucianisme. Va coeditar l'obra *Well-Being in Contemporary Society* (Springer, 2015).

Judith Simon

Judith Simon és llicenciada en Psicologia per la Freie Universität Berlin i Doctora en Filosofia per la Universitat de Viena. Actualment, és professora d'Ètica de la Informació i la Tecnologia a la Universitat d'Hamburg. També és part dels consells de redacció de revistes com *Philosophy and Technology* (Springer) i *Big Data & Society* (Sage), així com membre del Consell d'Ètica alemany. Ha estat becària visitant a la Universitat d'Stanford i investigadora convidada a Trento, a Ljubljana i a l'Institut de Recerca en Intel·ligència Artificial (CSIC-IIIA) de Barcelona. Els seus principals àmbits de recerca són els problemes epistemològics i ètics entorn de les tecnologies de la informació, la comunicació i la computació, com l'ètica informàtica, i les responsabilitats d'usuaris i dissenyadors de les TIC. El 2013 va rebre el premi Herbert A. Simon de la *International Association for Computing and Philosophy* (IACAP).