PERSPECTIVAS ÉTICAS APLICADAS SOBRE EL CAMPO DE LA INTELIGENCIA ARTIFICIAL

ANTONIO LUIS TERRONES RODRÍGUEZ
Universitat de València

RESUMEN: El desarrollo de la inteligencia artificial (IA) llevará a cabo importantes transformaciones en diversas esferas de despliegue de la acción humana. La introducción de los intelectos sintéticos en el universo vital del ser humano suscita una serie de problemáticas que deben ser abordadas desde la filosofía moral. En ese sentido, hay una serie de planteamientos éticos que se encuentran implícitos en los postulados de determinados teóricos de la IA y en este trabajo se identifican algunos. Esta identificación y posterior categorización dentro de diversas corrientes éticas tiene como finalidad un esclarecimiento de las iniciativas que desde la filosofía moral se han llevado a cabo para analizar una tecnología caracterizada por la ambivalencia.

PALABRAS CLAVE: Inteligencia artificial; filosofía moral; ética; sociedad tecnológica.

Ethical perspectives applied to the field of artificial intelligence

ABSTRACT: The development of artificial intelligence (AI) will bring about significant transformations in various spheres of human action. The introduction of synthetic intellects into the human lifeblood raises a series of issues that must be addressed from the perspective of moral philosophy. In this regard, several ethical approaches are implicit in the postulates of certain AI theorists, and some are identified in this paper. This identification and subsequent categorization within various ethical currents aims to clarify the initiatives that have been undertaken within moral philosophy to analyze a technology characterized by ambivalence.

KEY WORDS: Artificial intelligence; Moral philosophy; Ethics; Technological society.

Introducción

> La mirada tiene algo de extraño, de paradójico: la total facilidad de mirar contrasta con la dificultad de mirar bien. Si hay luz, con solo abrir los ojos se nos aparecen las cosas que nos rodean, pero en cambio hay que prestar atención, fijarse bien, para darse cuenta de según qué aspectos de la realidad y, sobre todo, para percibir las cosas de otra manera.
>
> (Esquirol, 2006: 14)

¿Deben ser aceptadas todas las oportunidades que ofrece la inteligencia artificial (IA), incluso si implican una transformación de nuestros límites morales? ¿Es posible introducir criterios éticos en los intelectos sintéticos para fortalecer la normatividad? ¿Puede orientarse el desempeño de los sistemas artificiales hacia el cultivo de virtudes? ¿Hasta qué punto puede la IA integrar valores universales en su programación? Estas son algunas de las cuestiones éticas que plantean una serie de teóricos de la IA para estimular la reflexión en el seno de la sociedad tecnologizada. A este respecto, el objetivo del presente trabajo consiste en la exposición, análisis y categorización de un conjunto de elementos teóricos que tienen su origen en diversas tradiciones éticas.

doi: 10.14422/pen.v81.i316.y2025.004

Las figuras científicas seleccionadas para analizar en este trabajo son Raymond Kurzweil, Nick Bostrom, Shannon Vallor y Bill Hibbard. Durante la última década la reflexión sobre la aplicación de la ética al campo de la IA se ha convertido en una de las principales tareas de la ética aplicada a la tecnología. Ese interés ha experimentado un incremento debido a los importantes avances y a la profundización de la complejidad de un fenómeno tecnológico que está cada vez más presente en nuestras vidas. En la medida que son adoptadas las nuevas oportunidades que brindan los sistemas artificiales, surgen un conjunto de problemáticas éticas que precisan ser examinadas. Las personalidades anteriormente destacadas han asumido esa tarea y desde la óptica del hedonismo, el utilitarismo, el pensamiento kantiano, la ética de las virtudes aristotélica y la teoría de la justicia de John Rawls han elaborado unos planteamientos que permiten trasladar la reflexividad moral a un ámbito profundamente tecnificado como es el de la IA.

1. Raymond Kurzweil: utilitarismo hedonista como impulso tecnológico

Raymond Kurzweil, nacido en New York en 1948, es un importante inventor, pensador, futurista y director de ingeniería en Google desde el año 2012. Revistas como *Forbes* o *Inc* lo han considerado «la máquina de pensar suprema» o «el legítimo heredero de Thomas Edison». A Kurzweil se le atribuyen algunos inventos como el primer sintetizador de voz para ciegos, el primer escáner CCD, el primer sintetizador de música capaz de recrear numerosos instrumentos, etc. Ha recibido múltiples premios y entre sus obras más importantes destacan *La singularidad está cerca* (2016) y *Cómo crear una mente* (2016). Durante las últimas décadas ha realizado importantes predicciones en el ámbito de la IA. El análisis de los planteamientos teóricos de Kurzweil permite afirmar que sus posiciones son afines al utilitarismo, una doctrina ética fundamentada en tradiciones como el hedonismo.

La búsqueda de la felicidad ha constituido un objeto de estudio esencial en la filosofía desde la Antigüedad. Aristóteles es una de las figuras más destacadas de la antigua Grecia que articula una reflexión ética orientada a la exploración del fin último a partir del concepto *eudaemonia*, traducido como felicidad. Se trata de una ética de carácter teleológico, puesto que se encuentra orientada hacia un *telos* o finalidad consistente en alcanzar la felicidad. Para el estagirita los placeres y la riqueza no suministran la verdadera felicidad, ya que esta se origina en la realización de aquello que es propio del hombre. De este modo, la vida contemplativa, la actividad propia del sabio, es para Aristóteles el medio más adecuado para alcanzar la verdadera felicidad. Sin embargo, más tarde, esta teorización de la felicidad experimenta una serie de cambios con el surgimiento del hedonismo y el establecimiento del epicureísmo como una escuela filosófica que centra su propuesta en el placer y el dolor como epicentros de la actividad moral.

El eje fundamental sobre el que circula la propuesta hedonista se encuentra en el placer y el dolor como criterios esenciales para el discernimiento moral. La ausencia de placer implica la búsqueda de satisfacción de una ausencia por medio de la apetencia del deseo. En la *Carta a Meneceo*, Epicuro comparte su propuesta filosófica sobre el placer, afirmando que el deseo de todo humano consiste en el disfrute del placer como un medio para la felicidad, constituyendo condiciones indispensables la salud corporal y la serenidad anímica, apreciando la ataraxia como la ausencia de toda perturbación. El reconocimiento del dolor y el temor, simbolizado principalmente por la muerte, representa una motivación para el alcance del placer como un medio sanador. En este sentido, la búsqueda del placer es para el hedonismo una causa moralmente razonable, debido a que supone el medio más adecuado para alcanzar la felicidad, con tal de situarse en la estela de la ataraxia. Así pues, el hedonismo propone un principio orientativo de las acciones morales que procura asegurar la salud en el cuerpo y tranquilidad en el alma, ofreciendo, de ese modo, un sustrato filosófico para tradiciones posteriores como el utilitarismo.

Los postulados teóricos de Jeremy Bentham y John Stuart Mill en el terreno de la filosofía moral constituyen una de las principales contribuciones del pensamiento inglés a la historia de las tradiciones éticas. Los dos grandes principios del utilitarismo pueden ser resumidos en las siguientes afirmaciones, según Esperanza Guisán:

a) El valor más importante es la felicidad a nivel individual;
b) y también el bienestar colectivo, vinculado a una utilidad generalizada, pues ha supuesto un importante objeto de reflexión en la historia de la filosofía y en la organización de gobiernos y políticas (2013: 274).

A este respecto existen vasos comunicantes entre el epicureísmo y el utilitarismo de Bentham y Mill en torno a la consideración del placer y la felicidad como dos máximas morales esenciales para guiar la acción humana. Por tanto, a continuación, argumentaré que el planteamiento de Kurzweil integra los principales elementos del utilitarismo hedonista para justificar moralmente la aceptabilidad de un alto nivel de tecnologización de la vida.

La idea fundamental que posibilita la identificación del planteamiento de Kurzweil como un postulado hedonista y utilitarista, es su defensa de la tecnología más avanzada, en particular de la IA, como un mecanismo para desafiar la muerte y el envejecimiento. En este caso, la muerte y el envejecimiento representan el dolor al que hacía referencia Epicuro, así como la constante preocupación por la ausencia de ataraxia. En este sentido, la biotecnología avanzada puede contribuir a la satisfacción del deseo consistente en el desafío a la muerte y el envejecimiento, alejando así del dolor que causan esos fenómenos ligados a la vida.

En este contexto se identifica el dolor con la sensación experimentada por el envejecimiento y la pérdida de funcionalidad, así como por la constatación de los límites impuestos por la condición de mortalidad. En particular, Kurzweil

manifiesta una notable preocupación por las enfermedades ligadas al sistema digestivo y al metabolismo, y apela al incremento de la presencia de dispositivos tecnológicos como los nanorobots para facilitar la eliminación de grasas y la absorción de nutrientes beneficiosos para el organismo, contribuyendo, de ese modo, a la felicidad del ser humano.

> En esa etapa de desarrollo tecnológico podremos comer lo que queramos, es decir, cualquier cosa que nos proporcione placer gastronómico. Exploraremos las artes culinarias para descubrir sabores, texturas y aromas, y además poseeremos un fluido óptimo de nutrientes en nuestro torrente sanguíneo […] En último término, no tendremos que preocuparnos de llevar prendas especiales o de contar con recursos nutricionales explícitos (2016: 349).

El progreso tecnológico ofrece un conjunto de oportunidades moralmente aceptables debido a que permiten mejorar las condiciones de vida, reducir los niveles de envejecimiento y alargar la vida. La felicidad, aspecto esencial del utilitarismo hedonista, es posible gracias a la integración de la inteligencia biológica con la inteligencia tecnológica. La gestión de la tecnología se representa en el pensamiento de Kurzweil una apuesta por la vida y un medio para la felicidad. La superación de las limitaciones y las dificultades ligadas al dolor inherente a la vida biológica, es ejecutada a través de un conjunto de dispositivos artificiales:

> En ese momento la longevidad de un fichero mental no dependerá de la continua viabilidad de ningún medio de hardware en particular (por ejemplo, la supervivencia de un cuerpo biológico y de un cerebro). En último término, los humanos basados en software se habrán expandido mucho más allá de las limitaciones humanas tal y como las conocemos hoy en día. Vivirán en la web proyectando sus cuerpos cuando quieran o lo necesiten, lo cual incluirá cuerpos virtuales en diferentes ámbitos de realidad virtual, cuerpos proyectados holográficamente, cuerpo proyectados mediante foglets y cuerpos físicos que contenga enjambres de nanorobots y de otras formas de nanotecnología (2016: 372).

En definitiva, la propuesta de Kurzweil se sitúa en la estela del utilitarismo hedonista, especialmente porque en el contexto de su obra identifica el dolor humano con el envejecimiento y la muerte ante el deseo de prolongación de la vida. Un deseo que es posible satisfacer a través de las oportunidades tecnológicas, y particularmente de la IA, a las que se les atribuyen propiedades facilitadoras para alcanzar la felicidad.

2. Nick Bostrom: la normatividad ética como guía de la acción

Nick Bostrom es un filósofo de origen sueco, director del Future of Humanity Institute y fundador de Humanity Plus (H+), la asociación internacional de transhumanismo. Es conocido por sus estudios sobre IA y mejoramiento humano, y entre sus obras más destacadas se encuentran *Superinteligencia,*

caminos, peligros, estrategias (2016) y *Mejoramiento humano* (2017), una obra coeditada junto a Julián Savulescu.

En un texto elaborado bajo el título *The ethics of artificial intelligence*, Bostrom realiza un análisis de algunos problemas éticos que han suscitado los sistemas que integran IA. La automatización de numerosos procesos se ha incrementado durante los últimos años, como es posible apreciar en la industria, el sector bancario, etc. Esto implica que el despliegue de la IA también ha motivado el surgimiento de importantes cuestionamientos éticos debido a las implicaciones morales relacionadas con la automatización. Los sesgos de automatización, la excesiva confianza en los resultados algorítmicos, los sesgos racistas o sexistas, entre otros, ponen de manifiesto la necesidad de transparentar los procesos de programación y despliegue de los dispositivos que incorporan IA. Es fundamental que las tecnologías de IA reúnan estándares de transparencia y, al mismo tiempo, puedan integrar mecanismos de seguridad para fortalecer los sistemas frente a posibles intentos de manipulación o alteración. Igualmente, los intelectos sintéticos inician complejas discusiones en torno al trazado de responsabilidad de los enjambres que configuran en el entramado de diseño, fabricación, programación y despliegue de las respectivas innovaciones.

Con motivo de la tecnologización de la sociedad, Bostrom propone una serie de criterios normativos para incorporar al entorno de la IA. Entre esos criterios pueden encontrarse los siguientes: responsabilidad, transparencia, auditabilidad, incorruptibilidad y predictibilidad, entre otros (2011: 3). Estos criterios normativos señalan un camino de fundamentación kantiano que se encuentra implícito en el trabajo del filósofo sueco. A este respecto, Bostrom comienza preguntándose sobre el estatus moral de las máquinas, analizando si pueden ser consideradas un fin en sí mismas, del mismo modo que las personas en el marco del pensamiento kantiano. En la *Fundamentación metafísica de las costumbres* Kant investiga la dignidad humana originada en la libertad moral y manifiestada en la tercera formulación del imperativo categórico: «obra de tal modo que te relaciones con la humanidad tanto en tu persona como en la de cualquier otro, siempre como un fin y nunca solo como un medio» (1999: 104). De esta máxima ética es posible derivar que la persona posee un estatus moral propio, sin embargo, lo que se pregunta Bostrom es si la IA presenta o no ese estatus.

En primera instancia no es posible atribuir estatus moral a una IA, una afirmación que comparten los expertos de este campo. Si bien el humano puede manipular, eliminar información, desconectar o destruir un sistema, sin que eso implique un daño experimentado en primera persona por los intelectos sintéticos, la discusión se origina cuando se trata de determinar qué aspectos deben reunirse para atribuir estatus moral a una máquina. Bostrom propone dos criterios morales (2011: 7):

a) Capacidad sintiente: tiene que ver con la capacidad de experimentar o sentir dolor y sufrir.

b) Sapiencia: hace referencia al conjunto de capacidades vinculadas con una inteligencia de un nivel superior, asociadas al autocontrol, la conciencia y a la razonabilidad.

Es conveniente que nos detengamos en estos criterios, ya que pueden existir casos en los que no se satisfaga una u otra condición. Por ejemplo, pese a que los humanos neonatos no cumplen con el criterio de sapiencia, no se les niega el reconocimiento de la dignidad. Lo mismo ocurre con los discapacitados mentales, que podrían prescindir de alguno de los criterios anteriormente mencionados. Si bien pueden existir particularidades que precisan ser estudiadas con detenimiento, para el objeto de estudio de este trabajo es importante seguir teniendo como referencia esos dos elementos propuestos por Bostrom.

Para sortear los posibles conflictos sobre la posesión de estatus moral en entidades no humanas, el filósofo sueco propone el principio de no discriminación de material y el principio de no discriminación de ontogenia (2011: 8-9). El principio de no discriminación de material señala que, si dos seres presentan la misma funcionalidad y experiencia consciente, y la diferencia estriba únicamente en el sustrato de su implementación, entonces ostentarían el mismo estatuto moral. Bostrom sostiene que para estar a favor de este principio podría emplearse un argumento similar al racismo. De la misma manera que no se considera a una persona de piel negra inferior por su color, tampoco debería afirmarse que un intelecto sintético construido con un material diferente al ser humano puede tener una consideración moral diferente. No obstante, esa consideración moral no se realiza debido a que el intelecto sintético pudiera tener la misma utilidad y funcionalidad que el ser humano, sino porque el material puede ser moralmente relevante en la medida que es diferente al ser humano. Este principio considera que no existen diferencias morales si el cerebro de un ser usa semiconductores o neurotransmisores.

En cuanto al principio de no discriminación de ontogenia, el sueco sostiene que, si dos entidades presentan una funcionalidad y conciencia experimental similar, y únicamente la diferencia estriba en el proceso de existencia, entonces, debería poseer el mismo estatus moral. La ontogenia describe el desarrollo de un organismo desde su gestación hasta su senescencia. Este principio se erige sobre la idea de que el estatuto moral no depende del linaje ni de la casta. De tal manera, que aquellas personas que han sido fruto de la fecundación in vitro, de la adopción, etc., no quedarían al margen de ser consideradas poseedoras de estatuto moral. Pues bien, el principio de no discriminación de ontogenia extiende esas ideas hasta los sistemas cognitivos artificiales, es decir, hasta la IA. Así pues, para Bostrom, el origen de una entidad no determina el estatuto moral.

En la obra *Superinteligencia, caminos, peligros, estrategias* el filósofo sueco considera que la determinación de los valores para incorporar en una superinteligencia es una tarea caracterizada por la complejidad. La cuestión de la

elección de valores presenta dificultades y, por ende, implica una fuerte exigencia en lo relativo a la buena elección para evitar equivocaciones que acarrearían graves consecuencias. La dificultad de la elección de valores estriba en el relativismo cultural característico del universo axiológico. Los valores suelen tener diferente consideración en función del ámbito cultural, razón por las que son propuestas diversas jerarquizaciones. Además, existe otra dificultad en el proceso de definición del objeto final de los sistemas superinteligentes, suscitando complejas problematizaciones filosóficas. Finalmente, es importante destacar la inexistencia de un consenso en torno a una teoría ética que cuente con validez universal.

En medio de este escenario de dificultades axiológicas Bostrom propone lo que él denomina «normatividad indirecta». El sueco afirma la incapacidad del ser humano para determinar un conjunto de valores que cuenten con validez universal. Además, señala que el dinamismo axiológico que acompaña a los cambios históricos dificulta el ejercicio de la determinación (2016: 210). Pese a que la humanidad ha experimentado un notable progreso moral en diversas áreas, aún existen dificultades para formular un conjunto de valores universalmente aceptables y duraderos en el tiempo. A este respecto, la normatividad indirecta formulada por el filósofo sueco radica en la justificación razonable para desarrollar una superinteligencia que cuente con la capacidad suficiente para seleccionar aquellos valores instrumentales que resulten moralmente más convenientes para los seres humanos (2016: 210-211). La superioridad cognitiva y epistémica de un sistema artificial de tal magnitud permitiría a la humanidad desatar este nudo gordiano.

3. Shannon Vallor: la posibilidad de cultivar las virtudes éticas

El trabajo de Shannon Vallor, profesora de la Universidad de Santa Clara en Silicon Valley, está motivado por el fenómeno de la automatización del campo profesional. Vallor realiza una reflexión de este fenómeno desde la óptica de la ética de la virtud. Si recordamos, Aristóteles dedica una de sus obras más importantes, la *Ética a Nicómaco*, a las virtudes, definiendo virtud como aquella disposición habitual a hacer el bien. Según el estagirita, en el hábito radica la formación de las virtudes, pues la repetición de determinadas acciones orientadas hacia el bien contribuye a que el ser humano vaya forjando un carácter virtuoso. A este respecto, Vallor encuentra en las profesiones la posibilidad de desarrollar hábitos virtuosos mediante el cultivo de la sabiduría práctica.

El cultivo de la virtud depende de los actos que la persona convierta en disposiciones habituales. Pero solo aquellas actividades que conducen al éxito, al bienestar, a ciertas habilidades orientadas a la adquisición de sabiduría, son las que podrían considerarse facilitadores de la virtuosidad. Aquí se encuentran, como señala Aristóteles, las actividades que están guiadas por una comprensión inteligente de la moral que es precisada por lo que manifiestan las situaciones

particulares (Vallor, 2015: 109). Este apunte de Vallor recuerda lo señalado por Aristóteles en el libro II de la *Ética a Nicómaco*, donde presenta el término medio como la capacidad para saber discernir lo prudente entre el exceso y el defecto. Así pues, en el marco del pensamiento aristotélico las actividades que contribuyen a la virtud surgen de la voluntad, se hacen de manera consciente y razonada, y además tienen que ver con una disposición habitual a hacer el bien. Si el progreso de la tecnología, impulsado en la actualidad por la IA, promueve actividades que comprometen negativamente o debilitan el cultivo de las habilidades, es posible afirmar que el carácter de los profesionales podría encontrar dificultades para encaminarse al término medio y, por tanto, al bien.

La pensadora estadounidense utiliza el concepto «descualificación» (en inglés, *deskilling*), procedente de la sociología, para referirse al fenómeno consistente en el desplazamiento de los trabajadores de su puesto de trabajo como resultado de la automatización y también a las consecuencias derivadas de ese desplazamiento, que implica la pérdida de valor y de mérito de los trabajadores humanos. Para Vallor, el fenómeno de la automatización del campo profesional ha resultado ambivalente debido a que ha conllevado consecuencias positivas y también negativas. Las consecuencias positivas hacen referencia a la liberación de tareas que tradicionalmente han resultado rutinarias, descargando al trabajador de la realización de determinadas actividades y potenciando su formación y capacitación. Sin embargo, la automatización también ha conllevado notables consecuencias económicas, sociales y culturales, hasta ahora desconocidas.

Las nuevas tecnologías han transformado un gran número de estructuras sociales, económicas y culturales, motivo por el que es importante desarrollar una reflexión crítica para analizar cómo están afectando a la formación del *ethos* en la actualidad. Más allá de análisis dogmáticos que fomenten el pánico irracional a la tecnología, urge afrontar este fenómeno con una reflexión crítica responsable que permita construir planteamientos a partir de argumentos sólidos. El impacto de la tecnología no debe caracterizarse por el desequilibrio, según Vallor, pues no tiene que ser únicamente un generador de impactos negativos en el carácter profesional. Así pues, una reflexión que gire en torno al fenómeno del progreso de la tecnología debe contar con aportaciones de la ética de la virtud, la psicología moral y un estudio empírico riguroso sobre las condiciones materiales y sociales que propician el florecimiento humano (2015: 112).

Para Vallor es deseable otorgar la importancia que merece a los efectos de la tecnología en el desarrollo de las habilidades morales, en vista de que esas habilidades son las que se encuentran directamente relacionadas con la formación de un carácter virtuoso. La restricción o limitación del uso de la tecnología no constituye una solución viable, según Vallor, debido a que es conveniente reconocer que la tecnología ha contribuido con importantes beneficios en numerosas áreas y ha propiciado un incremento del conocimiento tecnomoral (2015: 113). A este respecto, la investigadora identifica tres riesgos potenciales para el carácter moral: los vehículos aéreos no tripulados, en particular, los sistemas

armamentísticos autónomos; la tecnología de medios de comunicación; y los robots de cuidados.

Vallor es consciente del creciente desarrollo que ha experimentado la tecnología de IA en el campo militar, por ejemplo, en el caso del estadounidense X-47B o del británico Taranis. La tecnología militar se va perfeccionando con el paso del tiempo y está logrando importantes resultados en la lógica de funcionamiento militar. El éxito de las armas autónomas está motivando a las instituciones militares a aumentar su uso en los próximos años y a ampliar el espectro de acción. Vallor aplaude la renovación de la resolución de la Comisión Internacional de DDHH por parte de la Asamblea General de la Organización de Naciones Unidas (ONU) (2017), donde se realizan algunas consideraciones sobre la reglamentación en caso de conflictos militares. Para la norteamericana es fundamental la actualización de los reglamentos en materia de conflicto militar a la luz de la tecnologización. Es posible impulsar un desarrollo militar virtuoso mediante el cultivo de habilidades morales relativas al combate que se encuentren motivadas por el bien y el respeto a la legalidad internacional (2015: 114-115).

La alternativa propuesta por la norteamericana consiste en un aprovechamiento de las oportunidades que ofrece la IA para suministrar una mayor y mejor información al personal militar, con el propósito de disponer de mejores conocimientos para el desarrollo de su actividad profesional de un modo razonable. El distanciamiento con respecto al campo de batalla puede servir para que el personal militar no sufra traumas psicológicos y morales, como implica la utilización de los aviones no tripulados. La propuesta para lograr un equilibrio en el impacto de la IA en el campo militar, consiste en que los sistemas armamentísticos autónomos favorezcan el desarrollo de las habilidades morales de la mejor manera posible para el cultivo de la virtud. Como señala Vallor: «las tecnologías militares no son meramente herramientas para lograr objetivos prácticos; cada vez más, son el medio para definir el tipo de soldado que un soldado quiere ser y qué virtudes morales puede desarrollar a través de su servicio» (2015: 116).

Otro de los riegos que señala Vallor se encuentra en las tecnologías de la información y la comunicación (TIC). Numerosos estudios han puesto de relieve que la habilidad cognitiva para atender se ha visto mermada con el uso de estas tecnologías. La atención no es únicamente una habilidad cognitiva, sino también una habilidad moral. Las habilidades cognitivas se relacionan con la inteligencia moral de los seres humanos y también con otros tipos de intelectos, es decir, ayudan a saber cuándo prestar atención, a quién, durante cuánto tiempo y de qué manera, algo que es fundamental para la formación de un pensamiento crítico y una ciudadanía bien formada. El mal uso de las TIC socava la capacidad cognitiva y fomenta nuevas distracciones, disminuyendo la concentración. En ese sentido, la solución no se encuentra en una especie de combate contra la tecnología de la comunicación y la información. Más bien, la tarea podría consistir, según la norteamericana, en repensar el sentido de la

tecnología y la utilidad que se le atribuye, con el propósito de lograr la obtención de un mayor beneficio para el cultivo de un carácter virtuoso.

Por último, Vallor considera que otro de los riesgos puede originarse en el contexto de los robots dedicados a los cuidados (en inglés, *carebots*). En muchos países del mundo el crecimiento de la población adulta ha ido en aumento en las últimas décadas, despertando el interés en los desarrolladores de robots para proporcionar herramientas en un campo en el que pueden obtener una gran rentabilidad. Para la estadounidense, la introducción de los robots en el ámbito de los cuidados presenta importantes desafíos éticos, pues el cultivo humano de prácticas vinculadas con el cuidado, así como las virtudes morales que se derivan de esa actividad, podrían experimentar un preocupante debilitamiento (2015: 119).

No es fácil desarrollar la actividad de los cuidados, puesto que se encuentran implicados aspectos emocionales, físicos y financieros que se van configurando con el paso del tiempo y precisan de un importante componente de voluntariedad. La actividad de los cuidados integra un componente emocional que es fundamental para dar sentido a esa actividad. Dicha actividad es una de las que comúnmente hoy es reconocida como virtuosa y se va desarrollando en un estrecho vínculo con el otro, a partir del cual se va dando forma al carácter. El cultivo del carácter en los cuidados se va forjando de forma relacional con el otro condicionante. Así pues, Vallor sostiene que los desarrolladores de robots en el campo de los cuidados deben tener en cuenta todos estos aspectos y orientar su actividad hacia la creación de sistemas que satisfagan las necesidades morales de los cuidadores, también sus habilidades, pero sobre todo una mayor calidad en la actividad del cuidado considerando todos los matices de virtuosidad que subyacen en esa actividad.

Como ha podido comprobarse, Vallor ofrece interesantes alternativas a los tres riegos mencionados y todas ellas las construye desde la ética de la virtud. En términos generales, apuesta por estudiar qué posibilidades de cultivo de la virtud subyacen en las tecnologías que impulsa la IA y trata de potenciarlas para esclarecer los beneficios morales positivos que pueden ser obtenidos. No obstante, la filósofa considera que esto es necesario, pero no suficiente, y reivindica la necesidad de un profundo cambio de valores culturales en las sociedades tecnológicas para promover una nueva conciencia global en la que se asuma que el ser humano es a la misma vez artífice y artefacto tecnológico (2015: 122). Finalmente, insiste en un nuevo compromiso para fortalecer el cultivo de las virtudes en un tiempo marcado por altos niveles de tecnologización de las profesiones.

4. Bill Hibbard: el aprendizaje de valores basados en el principio de justicia universal

El estadounidense Bill Hibbard es un científico emérito del Space Science and Engineering Center de la Universidad de Wisconsin-Madison y también

colabora con el Machine Intelligence Research Institute. Entre sus escritos más importantes destacan *Super-Intelligent Machines* (2002), *Avoiding Unintended AI Behaviors* (2012) y *Ethical Artificial Intelligence* (2015). Sus investigaciones se centran en los problemas éticos que subyacen en el diseño de los sistemas de IA y cómo resolverlos.

Según Hibbard, las tres leyes de Isaac Asimov, aparecidas en su famosa obra *Yo robot,* contienen falacias y se caracterizan por la ambigüedad. Una de las soluciones que propone para resolver la ambigüedad de estas leyes consiste en la instrucción de los intelectos sintéticos por medio de la asignación de valores numéricos para cada resultado posible (2015: 10). A partir de ahí, el sistema podría usar esos valores numéricos para calcular qué decisión tomar. Sin embargo, Hibbard reconoce que el intelecto sintético podría dudar de los resultados de sus acciones, un hecho que implicaría la necesidad de realizar un cálculo de probabilidades. Esas probabilidades pueden servir para determinar los valores que espera se deriven de cada acción emprendida. La experiencia es un factor importante para el cálculo de probabilidades y también es un requisito fundamental para aquellos intelectos sintéticos que se centran en el aspecto probabilístico. En ese sentido, para Hibbard una IA que no cuente con experiencia previa no podría emprender acciones con un nivel de exigencia razonable (2015: 11).

Este modelo de asignación de valores numéricos a los resultados posibles se fundamenta en la idea de asociación de un beneficio con un resultado y, en cierta medida, ayuda a resolver las ambigüedades por medio de la instrucción. Para Hibbard, otro aspecto a considerar en la elección de los resultados consiste en la preferencia a partir de una función de utilidad para la obtención de beneficio. En este contexto, este científico destaca también la importancia de establecer una relación dialéctica con el entorno a partir de parámetros de acción y observación.

La estrategia de selección de acciones que impliquen una maximización del beneficio y una minimización del sufrimiento se encuentra en la línea del utilitarismo al que Hibbard critica. Para el norteamericano, tras los entresijos del utilitarismo se esconde una actitud permisiva, puesto que pueden ser permitidas acciones moralmente malas aunque presenten consecuencias beneficiosas. El utilitarismo es una ética normativa basada en reglas que pueden conducir a determinadas ambigüedades, ya que existen entornos que precisan la ruptura de algunas reglas. Una solución podría consistir en la introducción de un valor de utilidad para cada caso concreto en función de un historial de acciones y reglas que el agente transgrede (2015: 19).

Otra de las críticas que Hibbard formula contra el utilitarismo consiste en la ignorancia de la intencionalidad que se encuentra tras las acciones emprendidas. El comportamiento humano se suele conocer por medio de patrones y una función de utilidad que consista en la introducción de un historial de patrones de comportamiento de un agente permitiría contar con mayor información para conocer el comportamiento de ese agente con el entorno. No obstante, en lo relativo a la intencionalidad existen ciertas discusiones acerca

de si un intelecto sintético posee o carece de intencionalidad. Por ello, como señala Hibbard, tengan o no intencionalidad, la responsabilidad de aplicar la ética a esos sistemas depende de los desarrolladores humanos (2015:19).

La tercera objeción de Hibbard refiere a la imposibilidad de traducir aspectos de la vida humana a un lenguaje numérico y cuantitativo para introducirlos en los sistemas con IA. A este respecto el pensador señala que es importante la programación de preferencias en los sistemas de IA, con el propósito de no propiciar un sesgo subjetivo que pudiera escapar al control humano. Una vez expuestas las críticas y cuestionamientos que Hibbard formula al utilitarismo, el investigador insiste en el fortalecimiento de una perspectiva basada en las emociones para guiar el aprendizaje de los sistemas. Esta perspectiva facilitaría una programación algorítmica fundamentada en valores humanos (Hibbard, 2012: 113).

La propuesta relativa al aprendizaje de valores responde a la necesidad de evitar acciones intencionales que perjudiquen a los humanos. En este caso, la dificultad radica en la selección de unos valores lo suficientemente ricos como para enfrentarse a la gran variedad de situaciones que configuran la realidad y que normalmente se fundamentan en el plano subjetivo, un hecho que implica ambigüedad. Cuando los humanos se enfrentan a las diferentes situaciones que dan forma a la realidad asignan los valores que interfieren en las acciones que emprenden. No obstante, Hibbard menciona un estudio realizado por Luke Muehlhauser y Louie Helm, en el que los humanos no detallan con tanta facilidad ni precisión sus propios valores (Hibbard, 2015: 78). En este sentido, la complejidad axiológica pone de manifiesto la dificultad para trasladar un marco particular de valor al terreno algorítmico (2012: 113).

Como se acaba de mencionar, los sistemas de valores se caracterizan por la complejidad, aunque para Hibbard la dificultad no solo radica en la facilidad subjetiva para especificar los valores, sino también en la combinación de valores de varios individuos para incorporarlos a la IA. Es posible incorporar una suma de valores humanos a una ecuación matemática que posteriormente pueda formar parte del proceso de aprendizaje de un intelecto sintético. Para la incorporación de esa suma de valores a una ecuación matemática es importante un criterio de equilibrio de intereses entre varias personas, donde se conceda prioridad y un mayor valor a aquellos intereses que son compartidos y, de ese modo, se intente resolver el desacuerdo entre intereses por medio de la asignación de valores máximos y mínimos (Hibbard, 2015: 86-87). También sería conveniente que ciertos valores pasaran un filtro colectivo para establecer promedios de asignación, con la finalidad de poner solución a posibles inconsistencias.

En ocasiones, la determinación de los valores puede estar sometida a un cierto sesgo discriminador. La búsqueda de mecanismos para la combinación de valores de múltiples humanos y su integración a los sistemas de IA puede incentivar prácticas discriminatorias de determinados grupos sociales. Para tratar de corregir esa especie de discriminación por intereses, Hibbard recurre a la teoría de la justicia de John Rawls (2010). Rawls considera que en las

democracias liberales debe protegerse el principio de igualdad de oportunidades, con el fin de corregir las desigualdades económicas que puedan existir. Hibbard presenta su propuesta como alternativa al proyecto utilitarista que determina el valor de utilidad en base al cálculo de un promedio. A este respecto, el científico norteamericano reconoce que su propuesta para la determinación de valores entre múltiples humanos es posible que responda a la lógica utilitarista que él tanto crítica, motivo por el que rescata la propuesta de Rawls basada en el velo de ignorancia. El velo de la ignorancia de Rawls proporciona una matriz desde la que seleccionar valores múltiples para evitar así cualquier sesgo de intereses (2015: 87).

La revisión periódica constituye otro mecanismo para la introducción de valores en los intelectos sintéticos y consiste en un control periódico sobre los valores que han sido asignados para la IA. Esta revisión tiene la finalidad de llevar a cabo un proceso de adaptación al progreso moral que experimentan las sociedades en relación a su dinamismo histórico. Por tanto, para que los sistemas que integran IA no se encuentren en una situación de desfase, es importante que sean sometidos sus valores a una revisión periódica y a una posterior actualización (2015: 88- 89).

El rescate de Rawls propuesto por Hibbard supone una consideración de su planteamiento dentro del marco kantiano. Cuando se reivindica a Rawls para corregir los desfases utilitaristas en la selección de valores propuestos desde la multiplicidad de criterios humanos, se está también recogiendo el testigo del de Königsberg. Tras la inspiración rawlsiana en el planteamiento del científico norteamericano se encuentra el propósito de establecer unos principios de justicia universalmente aplicables y deducidos de una constitución común a todos los agentes morales. En definitiva, esta propuesta de aprendizaje de valores por medio de ecuaciones matemáticas que propone Hibbard es propiamente kantiana, debido a que se preocupa de dotarla de un carácter universal para corregir los desfases, a los que, según él, conduce el utilitarismo.

Conclusión

Si bien la IA suministra un gran número de oportunidades en diversas áreas, una serie de problemáticas derivadas de su diseño y aplicación han precisado sendos análisis éticos llevados a cabo por las personalidades científicas presentadas en este trabajo. Es indudable que estamos ante una tecnología que presenta una gran complejidad y que son los profesionales de este campo los que cuentan con los conocimientos técnicos necesarios. Sin embargo, en vista del despliegue que la IA experimenta en la sociedad y sus esferas, de las problemáticas originadas por su carácter ambivalente y del alcance de las disciplinas técnicas para examinar de un modo integral esta tecnología, es preciso establecer un diálogo con la filosofía moral.

La historia de la filosofía contiene un legado de tradiciones éticas desde las que es posible observar propositivamente los fenómenos de la vida moral. La aplicación de conocimientos procedentes del seno de la ética abre un abanico de oportunidades para analizar los límites y potencialidades de la IA. Aspectos como la utilidad, la noción de límite, la normatividad, los principios, las disposiciones virtuosas o la posibilidad de unos valores universales proceden de la historia de la filosofía moral y nos pueden permitir conocer con un mayor grado de responsabilidad epistémica la magnitud y las implicaciones de los sistemas que integran IA.

Bibliografía

Bostrom, N. y Yudkowsky, E. (2011). The Ethics Of Artificial Intelligence. En Ramsey W. y Frankish (eds.). *Cambridge Hankbook of Artificial Intelligence* (316-334). Cambridge: Cambridge University Press.

Bostrom, N. (2016). *Superinteligencia, caminos, peligros, estrategias*. Madrid: Teell Editorial.

Bostrom, N. (2017). *Asilomar AI Principles*. Disponible en: https://futureoflife.org/ai-principles/?cn-reloaded=1

Bostrom, N. y Sandberg, A. (2017). El mejoramiento como desafío práctico. En N. Bostrom y J. Savulescu (eds.). *Mejoramiento humano* (391-435). España: Teell Editorial.

Esquirol, J. (2006). *El respeto o la mirada atenta: una ética para la era de la ciencia y la tecnología*. Barcelona: Gedisa.

Guisan, E. (2013). El utilitarismo. En V. Camps, O. Guariglia y F. Salmerón (Eds.), *Concepciones de la ética* (269-296). Madrid: Trotta.

Hibbard, B. (2002). *Super-Intelligent Machines*. New York: Springer.

Hibbard, B. (2012). Avoiding Unintended AI Behaviors. En J. Bach, B. Goertzel y M. Iklé, (eds.). *Artificial General Intelligence*: 5th International Conference (107-116). New York: Springer, pp. 107-116.

Hibbard, B. (2015). *Ethical Artificial Intelligence*. Disponible en: https://arxiv.org/ftp/arxiv/papers/1411/1411.1373.pdf

Kant, E. (1999). *Fundamentación metafísica de las costumbres*. Barcelona: Ariel.

Kurzweil, R. (2016). *Cómo crear una mente. El secreto del pensamiento humano*. Berlín: Lola Books.

Kurzweil, R. (2016). *La singularidad está cerca. Cuando los seres humanos trascendamos la biología*. Berlín: Lola Books.

Rawls, J. (1997). *Teoría de la justicia*. México: Fondo de Cultura Económica.

Vallor, S. (2015). Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. *Philosophy and Technology*, 28, pp. 107-124.

Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford: Oxford University Press.

Universitat de València
antonio.terrones@uv.es

Antonio Luis Terrones Rodríguez

[Artículo aprobado para publicación en enero de 2020]