Ética para máquinas: Similitudes y diferencias entre la moral artificial y la moral humana*

Aníbal Monasterio Astobiza
UPV/EHU
anibal.monasterio@ehu.eus

Machine Ethics: Similarities and Differences between Artificial Morality and Human Morality

ISSN 1989-7022

Resumen: Sistemas artificiales autónomos como bots, chatbots, robots, asistentes virtuales e incluso animales artificiales (animats), cyborgs o plataformas computacionales cada vez más forman parte de nuestro entorno. Este reino máquina –la taxonomía de entidades digitales- crece muy rápido y puede tener consecuencias no-intencionadas no siempre buenas. La ética para máquinas (también conocida como ética computacional o moralidad artificial) es una nueva área de investigación que trata de implementar principios y preferencias morales en la toma de decisiones de máquinas y sistemas artificiales. En este artículo, comparo la ética para máquinas y la moral humana, subrayo sus similitudes y diferencias y también me pregunto si es plausible formalizar y reducir computacionalmente la moralidad en sistemas artificiales.

Palabras clave: ética para máquinas, moral humana, moral artificial, agencia moral, agentes morales artificiales

Abstract: Autonomous artificial systems such as bots, chatbots, robots, virtual assistants and even artificial animals (animats), cyborgs or computer platforms are increasingly part of our environment. This machine kingdom - the taxonomy of digital entities - grows very fast and can have unintended consequences not always good. Machine ethics (also known as computational ethics or artificial morality) is a new area of research that seeks to implement moral principles and preferences into the decision-making of machines and artificial systems. In this article, I compare both, machines ethics and human morality, and seek to understand their similarities and differences but also whether ethics is computable or it is plausible to code morality in artificial systems.

Keywords: machine ethics, human morality, artificial moral agents, moral agency

Aníbal Monasterio Astobiza: "Ética para máquinas: Similitudes y diferencias entre la moral artificial y la moral humana", en Mario Toboso, Manuel Aparicio y Daniel López: Debate: Ética, robótica y tecnologías asistenciales
***ILEMATA**, Revista Internacional de Éticas Aplicadas*, nº 30, 129-147

1. Introducción

Hasta donde yo sé, y de acuerdo con Dennett (1997), el primer homicidio cometido por un robot fue en 1981[1]. En la planta de Akashi -de la compañía Kawasaki Heavy Industries- un brazo robótico en mal estado empujó a un técnico que estaba reparándolo contra una fresadora de ruedas dentadas hasta aplastarlo. Desgracias en entornos laborales ocurren inevitablemente a pesar de nuestros esfuerzos en materia de seguridad y salud laboral. Como nos recuerda Dennett (1997, p 351), accidentes en los que una máquina (o sistema artificial) está implicada, también son bastante frecuentes.

¿Qué hizo este accidente especial? ¿Por qué se le considera el primer homicidio a manos de un robot? Asumiendo que caracterizamos los hechos como un homicidio cometido por una máquina, de estas preguntas se pueden derivar otras muchas preguntas inte-

* Agradezco el patrocinio del Gobierno Vasco para llevar a cabo una estancia posdoctoral en el Oxford-Uehiro Centre for Practical Ethics de la Universidad de Oxford y a esta última institución su cálida acogida. También agradezco al Center for Bioethics, Harvard Medical School, de la Universidad de Harvard por acogerme como Visiting Fellow en Global Health and Social Medicine. Este trabajo se ha realizado en el marco del proyecto de investigación KONTUZ!: "Responsabilidad causal de la comisión por omisión: Una dilucidación ético-jurídica de los problemas de la inacción indebida" (MINECO FFI2014-53926-R); el proyecto de investigación: "La constitución del sujeto en la interacción social: identidad, normas y sentido de la acción desde la perspectiva de la filosofía de la acción, la epistemología y la filosofía experimental" (FFI2015-67569-C2-2-P); el proyecto de investigación "Artificial Intelligence and Biotechnology of Moral Enhancement. Ethical Aspects (FFI2016-79000-P)"; EXTEND (Horizon 2020) e INBOTS (Horizon 2020). Email personal: anibalmastobiza@gmail.com

Received: 23/04/2019
Accepted:07/05/2019

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

resantes, entre ellas: ¿Quién es responsable de los crimines de un robot o máquina? ¿Es el robot/máquina responsable?, ¿es el ser humano?

El derecho como institución, la teoría legal, trabaja con ciertas clases de problemas éticos y legales bien definidos. Progresivamente, los robots y las máquinas están formando parte de la sociedad humana y será inevitable que regulemos sus acciones modificando las leyes existentes o creando nuevas leyes. Más que nada porque la presencia de máquinas en entornos compartidos con humanos traerá problemas (esperemos que no homicidios). Incoar leyes específicas para regular el comportamiento de los robots requerirá un amplio debate público entre todos los agentes sociales, incluido el público en general. Si un robot comete un crimen y esto es debido a un fallo en su diseño, no cabe duda que los responsables serán los diseñadores o la compañía que fabricó el robot. De hecho, los robots no serían las primeras entidades no-humanas en cometer delitos. Las corporaciones, empresas o personas jurídicas lo han hecho muchas veces en el pasado. Pero si el desarrollo de la Inteligencia Artificial (IA) hace posible crear máquinas (o sistemas artificiales) cada vez más autónomas e inteligentes, el derecho mercantil no será de justa aplicación ante un mal comportamiento y quizá haya que aplicar el derecho penal a los robots. Incluso puede que llegue a ser necesario otorgar un estatuto moral y legal de cuasi-persona a los robots[2]

Los recientes avances en robótica e IA prometen transformar la economía, la sociedad y nuestras vidas. Algunos están entusiasmados con los vehículos autónomos, drones mensajeros y otras manifestaciones de la IA, mientras que otros se sienten amenazados por la emergencia de máquinas y sistemas artificiales inteligentes que cada vez más rápido toman diversos roles en nuestras vidas. Los beneficios asociados con la IA y la progresiva digitalización de la sociedad son múltiples. Pero a medida que su aplicación se da en distintas dimensiones y facetas de nuestra vida no solo es necesaria la ética para guiar un uso correcto de las capacidades de las tecnologías emergentes, quizá sea una obligación moral crear sistemas artificiales con agencia moral, es decir, máquinas que tomen decisiones basadas en principios y preferencias morales.

Si el robot de la planta de Akashi de la compañía Kawasaki Heavy Industries hubiera tenido la capacidad de distinguir lo correcto de lo incorrecto -es decir, si hubiera tenido una moral artificial- ¿se hubiera producido el accidente? La ética para máquinas, también conocida como ética computacional o moralidad artificial, trata de responder a la pregunta: ¿cómo hacemos para que las máquinas o sistemas artificiales se comporten moralmente? Desde la ciencia ficción con las famosas tres leyes de la robótica de Asimov (1950), la posibilidad de que computadoras, máquinas, etc., puedan causar daño a los seres humanos ha sido un tópico o tema recurrente.

Pero esta posibilidad ya no es ciencia ficción. Armamento militar que opera sin supervisión humana puede atacar objetivos humanos (Singer 2009), algoritmos deciden la compra de activos y títulos en bolsa (Lewis 2014), seleccionan y contratan personal (Crawford y Whittaker 2016), discriminan negativamente (Larson et al. 2016) o espían y roban nuestros datos (Angwin 2015), bots (agentes de software autónomos) crean noticias falsas (fake news) en plataformas y redes sociales (Shao et al. 2018), etc.

Máquinas y distintas manifestaciones de IA pueden causar daño a la sociedad o individuos de formas hasta hace poco impensables. Y a medida que sigamos creando sofisticados sistemas

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

artificiales los potenciales riesgos de daño serán mayores. La ética para máquinas, por tanto, quiere crear máquinas morales. No simplemente aplicar nuestros principios y preferencias morales en la regulación y uso de la IA. Esta última área de estudios se conoce como "ética de la IA". La ética para máquinas quiere construir robots o sistemas artificiales que decidan o constriñan su comportamiento sobre la base de reglas éticas. Pero también como objetivo ulterior la ética para máquinas pretende utilizar los recursos computacionales para intentar crear nuevas formas de moralidad. Con la IA puede que la humanidad resuelva problemas y conflictos que por nosotros mismos no somos capaces de entender. Quizá la ética para máquinas o IA altamente creativa pueda desarrollar nuevos sistemas éticos totalmente diferentes, más eficientes y óptimos que las tradicionales escuelas y doctrinas filosóficas históricas creadas por los seres humanos.

En este artículo pretendo apuntar la dirección que la investigación en ética para máquinas está tomando, pero para ello veo necesario comparar la moral humana con la moral artificial (ética para máquinas), entender sus similitudes y diferencias, y hasta incluso comentar si es factible formalizar computacionalmente la moral.

En la sección 2 describiré (contando con los hallazgos en psicología moral, antropología y neurociencia, además de la teoría moral) cuáles son los consensos normativos acerca de qué atributos (o propiedades) son suficientes y necesarios para la agencia moral ya sea en seres humanos o máquinas. En la sección 3 basándome en los recursos empíricos de las ciencias humanas y en los recursos conceptuales de la filosofía moral y ética, contaré qué sabemos hasta la fecha sobre el desarrollo moral en los seres humanos para en la sección 4 hacer una comparativa -similitudes y diferencias- entre la moral artificial y la moral humana. Finalmente, en la sección 5 trataré la cuestión sobre si es plausible, dado el conocimiento de que disponemos, formalizar computacionalmente en sistemas artificiales la moralidad.

2. Requisitos para la agencia moral

Por agencia moral me refiero a las acciones voluntarias e intencionales derivadas de elecciones morales. En filosofía, tradicionalmente, la capacidad para realizar acciones se conoce como *agencia* (Anscombe 1957 y Davidson 1963). La agencia moral es, por consiguiente, actuar motivado por razones morales. Aunque existen algunas teorías que no demandan una teoría de la racionalidad o razones explicitas para que una acción sea el resultado de agencia moral: porque uno puede actuar racionalmente sin deliberación o actuar irracionalmente con deliberación (Arpaly 2004).

La agencia moral tiene dos dimensiones principales: dimensión activa y dimensión pasiva. En otras palabras, en la agencia moral nos encontramos con un agente que actúa motivado por una elección moral y un paciente que recibe los efectos de la acción. En este sentido, la agencia moral no solo concierne el mundo de los actos o acciones, sino también la percepción y observación de los mismos. Estas dos dimensiones bien diferenciadas de la agencia moral, la activa y la pasiva, imponen ciertos constreñimientos al comportamiento moral.

Por ejemplo, dentro de la dimensión activa, para que se dé agencia moral ha de asumirse autonomía y/o libre albedrío. Solo podrán ser considerados genuinos agentes morales quienes

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

estén libres de coacciones externas, sean verdaderamente autónomos, y, por supuesto, quienes ejerzan sus acciones en libertad[3]. El principal constreñimiento dentro de la dimensión pasiva es la atribución de agencia moral. Muy probablemente muchos animales no-humanos tienen una maquinaria cognitiva que les lleva a tomar decisiones más allá de respuestas pre-programadas ante ciertos estímulos del entorno, y se podría decir que actúan con ciertos grados de libertad. Pero creo que no muchas personas (seres humanos) les atribuyan agencia moral cuando actúan. La agencia moral es eminentemente una noción normativa. Solo aquellos agentes cuya conducta es libre, está motivada por requerimientos morales y razonan sobre esos requerimientos para aplicarlos al contexto adecuado pueden ser considerados agentes morales. Una persona tiene obligaciones (requerimientos) morales, supuestamente tiene libertad (véase la nota a pie de página número 3) y puede razonar sobre esos requerimientos dependiendo del contexto. Por consiguiente, es capaz de ejercer agencia moral. Un perro o un gato no tienen obligaciones (requerimientos) morales, sería argüible sus respectivos grados de libertad y también su uso motivado de razones. Por tanto, un gran número de animales no-humanos, por no decir todos, no son considerados genuinos agentes morales.

Entonces, la concepción estándar de agencia moral dice:

$\forall$ a, a es un agente moral $\leftrightarrow$ (si y solo si)
1) tiene obligaciones o requerimientos morales
2) tiene libertad (autonomía)
3) puede razonar sobre esos requerimientos

Esta concepción estándar de la agencia moral establece una conexión entre la acción y tener razones (obligaciones o requerimientos y capacidad para razonar sobre esos requerimientos). Así, de esta concepción se desprende el siguiente silogismo para definir la agencia moral de un agente:

Si X desea A, entonces la acción B le lleva a conseguir A y por tanto X realizará B.

Es evidente que además del problema de autonomía (autogobierno) y si el razonamiento es un proceso determinado o libre (distinción entre causas naturales y razones internas), existe el problema de identificar quién es un agente o no. La llegada de sistemas artificiales agudiza este problema y por supuesto en el caso de que llegáramos a construir máquinas autónomas con capacidad ética: ¿cuáles serían nuestras obligaciones o deberes para con ellas?, ¿deberíamos respetar su autonomía?, ¿concederles derechos?

Como veremos en la sección 4, la pregunta es: ¿Agentes morales artificiales contemporáneos cumplen estos requisitos? Pero antes veamos cómo el ser humano adquiere su agencia moral.

3. Desarrollo moral humano

Si podemos calificar la moral como una capacidad entonces no es óptima. Es una solución tosca para el problema social de la coordinación entre individuos y si la comparamos con otros conocidos sistemas de la mente humana nos damos cuenta claramente de su estado subóptimo desde un punto de vista evolutivo. Por ejemplo, el sistema auditivo si lo comparamos con

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

lo que es teóricamente posible vemos como es cercano a lo óptimo (Christensen-Dalsgaard y Carr 2008). El sistema visual también se acerca a un grado bastante óptimo (Artal et al. 2017). La vista es capaz de ver múltiples fotones en las condiciones idóneas. Sin embargo, nuestra capacidad moral no es óptima en comparación con lo que es teóricamente posible.

Si pudiéramos examinar y entender porqué la evolución ha desarrollado una capacidad o facultad como la moral de nivel subóptimo, la primera impresión sería de sorpresa. A lo largo de nuestra evolución como especie, la moral como capacidad para resolver problemas de coordinación entre individuos, es casi imprescindible. Los seres humanos hemos tomado un camino evolutivo que nos ha hecho seres gregarios y necesitamos de otros para sobrevivir. Desde la aparición del ser humano anatómicamente moderno hemos necesitado formar alianzas, cazar juntos, vivir juntos, etc., y como consecuencia inevitable entramos en conflicto.

Ahora bien, ¿qué mejor que una capacidad como la moral igual de óptima que la vista para percibir objetos, pero en su caso para percibir la confianza, emociones, intenciones… y otra información moralmente relevante y así evitar y resolver conflictos?

El quid de la cuestión es que la capacidad de la moral sí evolucionó para resolver problemas sociales de coordinación, pero no al nivel de optimización de otros sistemas o capacidades porque las demandas o presiones para que apareciera un sistema como la moral vinieron tarde en nuestra historia evolutiva. La mayor parte de nuestra historia evolutiva hemos vivido en grupos pequeños cuyos miembros estaban emparentados o relacionados genéticamente. Es con la progresiva vida en grupos más grandes donde cada vez más encontrábamos extraños y se convirtió en necesaria la aparición de normas y reglas para favorecer mejor la cooperación. De esta manera, entendiendo la moral como una facultad que se desarrolla como resultado de la presión de la vida en grupos grandes tenemos un desarrollo moral humano lleno de imperfecciones y anomalías, porque esta vida en grandes grupos con extraños apareció más tarde en la historia evolutiva del ser humano.

La moral no es nada misterioso ni sobrenatural. No nos ha sido regalada por una instancia trascendente o divina, sino que es simple y llanamente un conjunto de mecanismos biológicos y culturales para facilitar la cooperación (Greene 2013, Greene 2015, Curry 2016).

Mucho de lo que sabemos del desarrollo de la moral parte de la idea de que la interacción social es fundamental. La capacidad de la moral depende en buena parte de la interacción interpersonal. La interacción humana es en muchos contextos colaborativa. Una de estas interacciones, mejor dicho relación, a la que hay que prestar atención para poder entender la maduración y el desarrollo moral humano es la relación bebe-cuidador. Parte de nuestra supervivencia se debe a que somos capaces de formar una conexión socioemocional con nuestra madre, padre, cuidador… Las interacciones que se derivan de esta conexión primigenia entre el bebe y sus cuidadores son críticas para una verdadera inteligencia social y emocional y en muchos aspectos para la emergencia de la moralidad (Churchland 2017).

El contacto, físico piel con piel, entre cuidadores –principalmente madre- y bebe libera una hormona secretada por la hipófisis llamada oxitocina que crea un lazo de apego entre ambos. Este lazo de apego es tan fuerte que la madre ve a su bebe como una extensión de sí misma lo que le hace cuidar al bebe como si fuera ella misma. A su vez la oxitocina está asociada con la dopamina, un neurotrasmisor asociado con el procesamiento de recompensas. Por tanto,

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

el contacto y la interacción con el bebe se siente como placentera, lo cual a su vez hace que se libere más oxitocina y ésta más dopamina en un ciclo iterativo de retroalimentación.

La sociabilidad es por tanto un producto de la selección natural y la normatividad (moral) de los mecanismos de recompensa y de nuestra herencia mamífera. Durante la mayor parte del siglo XX la investigación sobre el desarrollo y adquisición moral estaba dominada por enfoques psicoanalíticos y conductistas donde la socialización parental era vista como fundamental y crítica. También durante el siglo XX, y bebiendo de la tradición racionalista kantiana, se construyeron las principales teorías empíricas del desarrollo moral humano (Kohlberg 1969 y Piaget 1932). Fue Piaget, pionero en el estudio del desarrollo moral, quien se aproximó al desarrollo moral a través de las justificaciones y juicios morales que emitían los niños ante dilemas morales. Así, describió dos estadios del desarrollo moral. Antes de los 7 y 8 años de edad los niños en este primer estadio de desarrollo siguen estrictamente las reglas dictadas por la autoridad (los adultos). Estas reglas las aceptan como válidas, se recompensa a quien las cumple y castiga a quien las transgrede. En este estadio solo se tienen en cuenta las consecuencias de las acciones y no las intenciones detrás de las acciones.

A partir de los 11 y 12 años se alcanza el segundo estadio de desarrollo moral. Es en este estadio donde los niños consideran que las reglas son modificables y son fruto del acuerdo social. También descubren que los adultos son algunas veces injustos en sus castigos. En este estadio, además de las consecuencias de las acciones, los niños también tienen en cuenta las intenciones de los agentes.

Por su parte, y siguiendo el trabajo de Piaget, Kohlberg también describió el desarrollo de los juicios morales siguiendo un proceso gradual con tres etapas diferenciadas. Estas tres etapas son: razonamiento preconvencional, razonamiento convencional y razonamiento posconvencional. El razonamiento preconvencional es el estadio más bajo caracterizado por un razonamiento moral heterónomo. Es decir, su fuente es externa con castigos y recompensas. Es de esta manera como los niños controlan su comportamiento. El razonamiento convencional es el estadio intermedio y es donde el niño empieza a internalizar las normas impuestas por otros (principalmente sus padres) y el entorno social (normas sociales). Finalmente, el razonamiento posconvencional es el estadio más alto donde las normas y valores se han internalizado y ya no se sigue las normas de otros y se reconoce que hay distintos juicios y justificaciones morales según el contexto.

Cada uno de estos tres tipos de razonamiento se subdivide en dos estadios, resultando en un total de seis estadios. El primero está basado en el castigo y la obediencia. El pensamiento moral se basa en el miedo. El segundo es instrumental e individual donde el razonamiento moral se basa en las recompensas y los intereses personales. El tercero es de expectativas mutuas. Los juicios morales se basan en valores personales aceptados, el cuidado y la lealtad a otros. El cuarto estadio se centra en el mantenimiento del orden social. La justicia, la responsabilidad y las normas se consideran las bases del razonamiento moral. El quinto estadio reconoce los derechos de cada individuo y la existencia de un contrato social. Valores como la justicia y la libertad son fundamentales. En el sexto estadio se reconocen los principios éticos universales. El individuo basa su razonamiento en principios compartidos y asumidos por todos.

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

Sin embargo, aunque los trabajos de Piaget y Kohlberg nos han arrojado luz sobre el desarrollo moral humano, la investigación reciente en distintas disciplinas de las ciencias sociales y ciencias cognitivas ha realizado una revisión completa de sus teorías y principalmente criticado sus premisas racionalistas basadas en Kant. Nucci y Turiel (1978) fueron los primeros en mostrar como los niños de entre 7 y 8 años distinguen entre transgresiones sociales y transgresiones morales, algo que contradice claramente las teorías de Piaget y Kohlberg, respectivamente, dado que según estos últimos los niños de entre 7 y 8 solo siguen normas morales dictadas por la autoridad (adultos, padres o cuidadores). Aún más, los niños siguen considerando que algo está mal aunque el mayor (autoridad) no lo vea y consideran que está mal aunque ocurra en distintos espacios (la escuela, la calle, en casa…).

Pero la principal crítica a los trabajos de Piaget y Kohlberg es una enmienda a las premisas racionalistas de origen kantiano de sus modelos. La reciente investigación en ciencias sociales y ciencias cognitivas afirma que parte de la competencia moral, emitir juicios, descansa en intuiciones socioemocionales (Haidt 2001). Haidt (2001, p. 822 y ss.) parte de un fenómeno llamado "atontamiento moral" (*moral doumbfounding*) en el que los adultos no saben articular y/o justificar porque tienen las intuiciones morales que tienen.

Esto contrasta con el racionalismo kantiano que subyace a los modelos de Piaget y Kohlberg donde reglas y principios son accesibles conscientemente y se pueden verbalizar tanto por niños como adultos. De acuerdo con Haidt, un evento moral genera una reacción afectiva-emocional que determina nuestros juicios morales. No obstante, en un intento de revitalización del racionalismo otros autores como Mikhail (2007) afirman que la moral consiste en una gramática moral que ante un evento moral identifica la estructura causal e intencional y que por tanto los juicios morales no emergen de intuiciones afectivas, sino que se basan en computaciones inconscientes de principios y reglas sobre la base de esa estructura causal e intencional del comportamiento de los agentes morales.

De esta tensión entre, por un lado, el intuicionismo moral de Haidt, y por el otro, la teoría de la gramática moral de Mikhail, en los últimos tiempos ha surgido una vía intermedia conocida como la teoría de los procesos duales de la moral: tanto la razón como la emoción de manera paralela, a veces en armonía, pero otras de manera contrapuesta y en conflicto, generan los juicios morales (Cushman et al. 2010, Greene et al. 2009).

En la actualidad, la investigación sobre el desarrollo moral humano se encuadra en una perspectiva evolucionista, entendiendo que la moralidad no es exclusiva de los seres humanos, sino que se comparte con nuestros parientes no humanos más cercanos como los grandes simios (Tomasello 2018, 2019). Desde esta perspectiva la cooperación favorece el bienestar de todos los individuos dentro de un grupo porque permite la interdependencia mutua.

Así, podemos recapitular algunos ingredientes básicos (no es una lista exhaustiva) de la capacidad de la moral humana dada la particular trayectoria evolutiva que hemos presentado:

- Selección de parentesco (lazos fuertes de apego con nuestros familiares).
- Teoría de la mente (razonar sobre otros como agentes con estados mentales propios).
- Altruismo/cooperación (extender el comportamiento prosocial y emociones aparejadas a extraños, ya sea por reciprocidad directa o indirecta).

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

Evidentemente, estos mecanismos o ingredientes que presento contribuyen a una narrativa evolutiva de carácter especulativo, pero basada en evidencias, de la moralidad humana, que no incluye todos los factores que posibilitan la moral. Muchos autores creen que el lenguaje es fundamental (Costa et al. 2014), otros que es necesario para la moral el pensamiento contrafáctico (Byrne 2017), la imaginación (Narvaez, y Mrkva 2014) o el razonamiento causal (Lagnado y Gerstenberg 2017), etc.

Otros autores enfatizan procesos de reciprocidad (Alexander 1987/2017), detección del engaño (Cosmides y Tooby 2004), empatía y simpatía (De Waal 2016), altruismo (Kitcher 2011). Esto es un indicativo de que no existe un consenso sobre la naturaleza de la moralidad. No obstante, como hemos ido viendo vamos teniendo una idea más empírica y adecuada del desarrollo y origen de la moral humana, a pesar de las dificultades. En la siguiente sección veremos cómo de manera intencional se quiere crear moral artificial en máquinas.

4. Programa general de la ética para máquinas

Todo intento de crear inteligencia artificial general o IA de nivel humano -en definitiva, máquinas inteligentes- debe buscar entender cómo se generan propiedades de alto-nivel como la moralidad. ¿Una IA puede tener inteligencia nivel humano sin moral? Creo que no. Como es muy posible que no se puedan crear máquinas o sistemas artificiales inteligentes sin lenguaje, autoconsciencia, emociones… Sin embargo, la posibilidad de una superinteligencia o gran inteligencia no implica deseos o estados apetitivos. Nuestros actuales sistemas artificiales son relativamente inteligentes, pero no quieren nada y no tienen desarrollado ningún sentido moral. Lo cual implica que se puede ser inteligente y no tener estados conativos. Pero creo que para facultades como la creatividad, moral y otros fenómenos es necesaria la autoconsciencia. Pero centrémonos en el programa general del campo interdisciplinar conocido como "ética para máquinas", que trata de imbuir a agentes morales artificiales la capacidad para tomar decisiones morales.

De las múltiples técnicas que se han desarrollado en los últimos años en el campo de la IA para realizar tareas de reconocimiento de patrones, análisis de imágenes, traducción automatizada, clasificación, predicción, etc., existe una estrategia en boga que es el uso de juegos (e.g. backgammon, damas, ajedrez, juego Go, Atari retro juegos, poker…) concebidos como entornos de entrenamiento para entrenar algoritmos y que mediante aprendizaje condicionado (aprendizaje a través de la práctica) se busca que sean capaces de realizar ciertas tareas que son propias de la inteligencia humana. Una de las compañías subsidiarias de Alphabet (Google) es DeepMind y utiliza juegos como entornos de aprendizaje para sistemas de IA.

Los juegos permiten crear entornos de multi-agentes que recrean las interacciones personales de las sociedades humanas. Las interacciones multi-agente son esenciales en la vida humana. Sin la interacción entre multi-agentes no serían posibles los gobiernos, organizaciones, grupos, ni los mercados económicos. A las interacciones multi-agentes subyace la colaboración y cooperación humana que, como hemos visto más arriba, es parte central de la moralidad. Los juegos recrean interacciones multi-agentes donde como en las sociedades humanas el comportamiento de un agente depende de los actos de otros agentes. En algunos juegos, como el ajedrez, toda la información (reglas, movimientos de las piezas…) es conocida por to-

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

dos los jugadores. Pero hay juegos, como la vida misma, donde la información es incompleta. Por ejemplo, en la vida real de interacción multi-agente cuando salimos a la calle no sabemos las intenciones de otras personas y gracias a pistas verbales y no-verbales inferimos y deducimos las intenciones detrás de sus actos (teoría de la mente).

En juegos de información incompleta se recrea esta característica de la vida social humana donde no siempre hay información completa. Con los juegos que simulan las interacciones multi-agente de la vida humana permites a agentes de IA ser capaces de cooperar efectivamente con otros agentes artificiales y al mismo tiempo con seres humanos. Los juegos de mesa cuya información es completa y toda ella es trasmitida a sistemas de IA a través de pixels en la pantalla son dominios estructurados. Si entrenas a un sistema de IA en estos dominios estructurados no tendrá versatilidad y todo lo que aprenda no podrá transferirlo o generalizarlo a otro dominio.

Por ejemplo, Alpha Go el programa informático de IA que batió al campeón del mundo en el juego de mesa Go, Lee Sedol, es mejor que un humano en este juego, pero toda su capacidad de procesamiento y poder computacional no se puede aplicar para jugar al parchís. Le sacas de las reglas del juego Go y es estúpido o por poner una imagen más intuitiva este programa si se está quemando la sala, aunque juegue mejor al Go que cualquier ser humano, no pondrá los medios más sencillos para salir de la sala ardiendo algo que no ocurriría con un niño de 4 años.

Sin embargo, los juegos de información incompleta o donde hay interacciones multi-agentes son una vía para intentar conseguir Inteligencia Artificial General o inteligencia nivel humano y, para nuestro propósito, moralidad. En esta línea, investigadores de DeepMind han conseguido avanzar un paso más en la búsqueda de inteligencia nivel humano en sistemas artificiales con el juego Hanabi, que es juego cooperativo de dos o más jugadores proveyendo al sistema artificial de capacidad de razonamiento sobre las creencias e intenciones de otros agentes (Bard et al. 2019).

En otra investigación llevada a cabo por la misma compañía (Rabinowitz et al. 2018), diseñaron una red neuronal para razonar sobre estados mentales (e.g. ToMnet) que pasa exitosamente el test de Sally-Anne (Baron Cohen et al. 1985) y por tanto es capaz de modelar y reconocer que otros tienen falsas creencias sobre el mundo.

Con el uso de algoritmos de aprendizaje por refuerzo autónomo y juegos concebidos como entornos de entrenamiento de los algoritmos para que aprendan con la práctica, compañías como DeepMind están aproximándose a la tarea de conseguir inteligencia nivel humano con todo lo que conlleva, incluida la competencia moral, de una manera orgánica, desde cero, o dicho de otra manera: desde un enfoque *abajo-arriba*. Pero también dentro de la ética para máquinas cuyo objetivo último es conseguir producir máquinas morales existe otro enfoque que busca programar reglas o meta-reglas de manera inicial para producir comportamiento moral. Este es el conocido como enfoque *arriba-abajo*. Así, tenemos dos enfoques principales para responder a la pregunta: ¿Cómo codificar o programar la moral en las máquinas? Por un lado, el enfoque *abajo-arriba*, y por otro lado, el enfoque *arriba-abajo*.

El enfoque *arriba-abajo* consiste en la implementación de habilidades morales en máquinas a partir de principios, reglas, etc., impuestas a partir de doctrinas éticas u otras fuentes como la religión o la literatura (Wallach, Allen y Smit 2007, p. 573). Desde este enfoque el ingeniero

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

que diseña la arquitectura de control del sistema artificial con agencia moral, introduce una serie de reglas o principios que constriñen el comportamiento. Dos teorías éticas que tienen una correspondencia directa con el enfoque *arriba-abajo* son: el consecuencialismo y el deontologismo.

Estas dos teorías éticas tienen una correspondencia directa con el enfoque *arriba-abajo* porque son susceptibles de formalizarse lógicamente. Para el caso del consecuencialismo un acto es moral si maximiza la "utilidad" (felicidad) del mayor número de personas (Bentham 1789/1996). La heurística de implementación del consecuencialismo en un software para que guie las decisiones de un agente moral artificial bien podría ser la siguiente:

> ∑ (intensidad × duración × probabilidad)

Esta heurística se lee: el sumatorio de parámetros como la intensidad, duración y probabilidad de una función que maximice el bienestar del agente artificial y otros establece la felicidad (utilidad) general. De esta manera, si un agente moral artificial cuya arquitectura de control ha sido diseñada desde las premisas del enfoque *arriba-abajo* y en concreto desde la doctrina ética del consecuencialismo, entendería que los actos morales son solo aquellos que tienen las mejores consecuencias para el individuo y la sociedad. Un algoritmo computacional es capaz de procesar múltiples alternativas, más que un ser humano. Un sistema artificial con competencia moral basado en un enfoque *arriba-abajo* desde las premisas del consecuencialismo es tecnológicamente viable. Instalar sensores para crear un modelo interno de la situación o entorno, generar posibles cursos de acción, predecir las consecuencias de las acciones y al mismo tiempo evaluar la situación en términos de utilidad es razonablemente realizable en sistemas artificiales.

Por su parte, el deontologismo también es una doctrina ética que casa bien con el enfoque *arriba-abajo* dentro del intento de codificar y programar habilidades morales en las máquinas. El deontologismo entiende que un acto es moral si está basado en una máxima, principio o regla universal. Un modelo de sistema artificial basado en el deontologismo se puede nutrir de una de las reglas más conocidas: el imperativo categórico de Kant (1785/2006). Este imperativo categórico dice así:

> Actúa sólo según aquella máxima por la que se puede, al mismo tiempo, desear que se convierta en una ley universal

La heurística de implementación en un software basado en el deontologismo diseñado para guiar la conducta de un sistema artificial asume que si una contradicción emerge la acción es rechazada, si hay armonía y acuerdo se acepta. De aquí parte la posibilidad de tomar decisiones morales siguiendo un modelo deontologista. Una formalización de esta heurística podría ser la siguiente:

$$obl(\alpha, \varphi) \ \neg per(\alpha, \neg\varphi)) \longrightarrow goal(\alpha, \varphi)$$

donde la obl (obligación) de α para hacer φ conduce a su realización si y solo si es per (permisible).

Un sistema artificial que bajo un enfoque *arriba-abajo* implemente una moral deontologista es a día de hoy tecnológicamente viable. Un software escrito en lógica deóntica con descripción de alto-nivel de operadores lógicos y categorías como: prohibido, permitido, necesario,

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

obligatorio... es posible (e.g. MedEthEx de Anderson, Anderson y Armen 2006 o Bringsjord. y Taylor 2014).

El enfoque *abajo-arriba* es una estrategia de implementación de habilidades morales en máquinas opuesta al enfoque *arriba-abajo*. En lugar de instalar en la programación una serie de principios o reglas morales que sigan las prerrogativas de una determinada doctrina ética, el enfoque *abajo-arriba* se limita a crear las condiciones sin estructura previa para que se aprenda de cero qué reglas morales son las más convenientes en un determinado contexto a través de la interacción con otros agentes. A este respecto, la ética de las virtudes como ética normativa que afirma la idea de que el carácter se puede desarrollar y por tanto es una teoría basada en el agente y no en la acción es una de las teorías que se corresponde directamente con el enfoque *abajo-arriba* (Abney 2014).

Para la ética de las virtudes un acto moral es el que hace un hombre/mujer de carácter bueno que da lugar a una conducta buena. Si se quiere desarrollar esta forma de ser que es la virtud y que da felicidad (*eudaimonia*) un sistema artificial debe implementar una arquitectura de control basada en redes neuronales que permita el aprendizaje versátil y flexible.

Además de estos dos enfoque genéricos (enfoque *arriba-abajo* y enfoque *abajo-arriba*) existen enfoques *híbridos* (Wallach et al. 2010) que combinan los aciertos de ambos enfoques. En el programa general de la ética para máquinas la misión y el objetivo principal es implementar razonamiento moral en sistemas artificiales y para ello no hay impedimento alguno por utilizar las técnicas disponibles que brinda la ciencia computacional. No obstante, es un error seleccionar una técnica favorita y una doctrina ética *ad hoc* e intentar que la teoría ética se adecúe a la técnica. Como bien muestran los enfoques vistos más arriba, no siempre una técnica computacional o enfoque es adecuado para toda teoría ética. También es importante realizar una cartografía general de lo que es posible dado el estado actual del arte en ética computacional o ética para máquinas. Como veremos en la sección de discusiones quizá no sea posible por el momento desarrollar un agente o sistema artificial moral con verdadera capacidad para el razonamiento y toma de decisión moral. Igual solo es posible restringir o constreñir el comportamiento de sistemas artificiales para que su conducta sea moralmente aceptable o deseable y que el sistema sea útil.

Por ejemplo, James H. Moor (2006) cree que no es posible crear agentes morales artificiales o utilizando su terminología "agentes éticos plenos" simplemente adhiriéndonos a argumentos filosóficos o a la investigación empírica. Es por ello que sugiere que dediquemos nuestros esfuerzos a desarrollar "agentes éticos explícitos" limitados (Moor 2006, p. 21).

Moor (2006) distingue entre varias formas de agentes morales artificiales: "agentes de impacto ético", "agentes éticos implícitos", "agentes éticos explícitos" y "agentes éticos plenos". Un agente de impacto ético es una máquina que ayuda a evitar una situación inmoral. Moor pone el ejemplo de los jinetes de camellos, que son niños pequeños tratados como esclavos. Es posible que podamos desarrollar robots que sustituyan a estos niños pequeños y entonces estaremos evitando una situación inmoral (Moor 2006, p. 19). Un agente ético implícito es una máquina programada para actuar éticamente o que evita el comportamiento inmoral porque sigue la programación de un diseñador que a su vez sigue principios éticos. Ejemplos son el piloto automático en aeronaves, la web de un banco, etc. Un agente ético explicito, por su parte, utiliza principios éticos explícitos para resolver dilemas morales. Por último, un

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

agente ético pleno puede emitir juicios éticos explícitos y es capaz de justificar dichos juicios (Moor 2006, p. 20).

Dentro del programa general de la ética para máquinas una distinción analítica acertada es la que proponen Dennis y Fisher (2018). Ellos distinguen: retos filosóficos y retos prácticos o técnicos. Los primeros retos se refieren a la dificultad de definir los valores, los principios éticos que representan dichos valores e identificar los mecanismos o procesos multinivel (neuronales, psicológicos, culturales…) implicados en la toma de decisiones morales. Por no decir que no hay ninguna teoría ética que recoja de manera suficiente y comprehensiva la complejidad del fenómeno moral o que no hay una teoría ética que agote la diversidad del fenómeno moral. Por retos prácticos se refieren a las cuestiones de cómo implementar el razonamiento ético en máquinas o sistemas artificiales.

Además de estos retos filosóficos y prácticos de indudable importancia, Gordon (2019) habla de otros retos metodológicos: codificar la ética en máquinas depende de cómo los programadores perciben en un primer momento qué es la ética y por tanto considera necesario que exista un diálogo entre programadores y filósofos.

5. Similitudes y diferencias entre la moral artificial y la moral humana

Hemos visto en distintas secciones lo que sabemos -de acuerdo a datos de la psicología moral, antropología y neurociencia- del desarrollo y origen de la moral humana y el intento general de codificar moral en sistemas artificiales (secciones 3 y 4, respectivamente). Ahora toca hacer un balance de las similitudes y diferencias entre la moral humana y la moral artificial.

Al final de la sección 2 se lanzaba la siguiente pregunta: ¿Agentes morales artificiales contemporáneos cumplen estos requisitos? Recapitulando lo dicho, la concepción estándar de agencia moral dice que para que exista agencia moral se han de dar los siguientes requisitos:

$\forall$ a, a es un agente moral $\leftrightarrow$ (si y solo si)
1) tiene obligaciones o requerimientos morales
2) tiene libertad (autonomía)
3) puede razonar sobre esos requerimientos

Dada esta concepción estándar de la agencia moral toca valorar si estos requisitos suficientes y necesarios de agencia moral se dan en sistemas artificiales contemporáneos. En la sección 5 trataré otra pregunta importante: ¿Cómo la competencia moral se puede implementar computacionalmente y si es plausible formalizar y reducir computacionalmente la moralidad en sistemas artificiales?

En la siguiente tabla vemos una comparativa a partir de la narrativa evolutiva de carácter especulativo, pero basado en evidencias, más arriba expuesta en la sección 3: propiedades, procesos o competencias de la capacidad moral humana (columna de la izquierda) con la implementación, instalación, codificación, emergencia o desarrollo de esas mismas propiedades, procesos o competencias morales, pero en sistemas artificiales (columna de la derecha).

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

MORAL HUMANA	MORAL ARTIFICIAL
simpatía/empatía	?
procesos cognitivos de atención conjunta	?
procesos sociales/cognitivos de agencia y cooperación	x
procesos de auto-regulación	?
emociones	?
capacidad lingüística	?
cognición social	?
razonamiento causal	x
razonamiento contrafáctico	?
razonamiento de sentido común	x
autoconsciencia	x
ayuda a parientes	x
lealtad al grupo	x
reciprocidad	x
respeto a los superiores	x
ser valiente	x
dividir recursos en disputa	x
respetar la propiedad privada	x

La tabla no quiere recoger -por otra parte, sería harto difícil hacer una revisión completa de la literatura de investigación en ética para máquinas- lo que investigadores o autores particulares, grupos de investigación, etc., han realizado conducente a la implementación de competencias morales en sistemas artificiales. La tabla se divide con una línea horizontal en dos partes. La parte de arriba está dedicada a identificar las competencias que he descrito como necesarias para dar lugar a una agencia moral de acuerdo con un relato hipotético de evolución y origen de la moral; y una segunda parte que siguiendo a Curry, Mullins y Whitehouse (2019) recoge formas especificas de comportamiento cooperativo que la moral predice.

Sin embargo, note el lector que en la columna de la derecha solo hay dos signos: "?" y "x". El primero es el signo de interrogación y viene a decir que aunque haya grupos de investigación, investigadores o autores que trabajan en trasladar las correspondientes competencias morales a máquinas no hay equivalencia artificial de dicha competencia o comportamiento similar a la contraparte humana. El segundo signo es "x" que denota desconocimiento, tanto porque no existen -o por lo menos no me consta a mí- grupos de investigación, investigadores o autores cuyo trabajo se dedique a la generación de dichas competencias o comportamientos en sistemas artificiales tanto por su dificultad empírica o teórica de poder hacerlo. Véase, para más información sobre la miríada de competencias asociadas con la moral humana y su posible implementación en arquitecturas computacionales y los problemas que ello acarrea, Malle y Sheutz (2019), Asaro (2016), Vanderelst y Winfield (2018).

A pesar de los esfuerzos de muchos filósofos de teorizar sobre la ética para máquinas y tecnólogos por crear máquinas con la capacidad de tomar de decisiones morales por sí mismos y, lo más importante, que no haya ninguna contradicción lógica de crear sistemas artificiales con agencia moral; como hemos visto en la tabla comparativa muchas de las características de la moralidad humana siguen sin tener una contraparte en la moral artificial. La moral humana

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

ha evolucionado y desarrollado una serie de competencias que dan lugar a la competencia o facultad de juzgar moralmente eventos, acciones o personas y los productos de esta competencia, normas, valores, etc., pueden cambiar. Hasta que las máquinas sean capaces de crear su propia cultura y entorno social desde el que desarrollar su propia moral, es muy probable que el enfoque *arriba-abajo*, caracterizado por ser los ingenieros y programadores quienes instalen o programen prerrogativas morales a sistemas artificiales, sea la única manera en la que máquinas tengan moral. Pero esto conduce a que nuestras preferencias cambien y las máquinas actúen en contra nuestra.

Bryson (2018) dice que no es inevitable la ética para máquinas. Al ser una decisión consciente e intencional del hombre crear máquinas con agencia moral quizá la elección más ética sea no crear máquinas morales. Que tengamos la capacidad tecnológica de crear máquinas con agencia moral no implica que debamos crearlas. No obstante, todavía sigue siendo un gran reto técnico y práctico crear ética para máquinas.

Como el propio Dennett (2019, p 41) dice, el mismo autor con el que empezábamos, "no necesitamos agentes artificiales conscientes (morales)". La creación y desarrollo de tecnología que transforme radicalmente la forma en la que nuestras habilidades, incluida la moral, son ejercitadas con otras personas o con otras máquinas puede tener un impacto en la forma en la que organizamos nuestras vidas y sociedades. En un mundo previo a la robótica y la IA, los aspectos esenciales de nuestra condición como la moralidad como cooperación, la interdependencia conjunta de unos con otros, la amistad, la intimidad, el sexo… eran facultades que servían instrumentalmente a la interacción y socialización humana. En un mundo donde las maquinas, cada vez más, pueblan nuestros entornos y median nuestras relaciones, esto puede afectar nuestra condición humana. La posibilidad de crear máquinas morales, robots compañeros y de asistencia, robots sexuales cambia nuestra moral, nuestro sentido de la amistad y nuestra sexualidad. Al querer antropomorfizar y conferir humanidad a máquinas artificiales corremos el riesgo de que las máquinas y la tecnología nos automaticen, roboticen o nos deshumanicen a nosotros. Como reconoce Christakis (2019), para bien o para mal, los robots y la IA cambiarán nuestras habilidades humanas como la amistad, sexualidad y hasta incluso nuestra moral.

6. Discusiones

El reino máquina -taxonomía de entidades digitales (Hernandez-Orallo 2017)-crece muy rápidamente y un número cada vez más grande de sistemas digitales y software alcanzan un nivel de agencia tal que pueden interactuar con nosotros. A medida que estos agentes artificiales toman decisiones que nos afectan, muchos creen que una forma de evitar consecuencias negativas es imbuir a estos sistemas artificiales de capacidades de razonamiento moral. La ética para máquinas, también conocida como ética computacional o moralidad artificial, es un campo interdisciplinar que busca imbuir de razonamiento y capacidad para toma de decisiones morales a sistemas artificiales. No obstante, vista la comparativa de similitudes y diferencias entre la moral artificial y la moral humana, la moralidad (y por ende la ética) sigue siendo una competencia que descansa exclusivamente en el dominio humano (Floridi y Sanders 2004, p. 374, Latorre 2019, Malle y Sheutz 2019).

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

Me toca ahora tratar una última cuestión, pero no por ello menos importante: ¿Es plausible formalizar y reducir computacionalmente la moralidad en sistemas artificiales? Muchos investigadores se preguntan sobre la forma en la que máquinas o robots deben tomar decisiones morales (Awad et al. 2018). Para ello testan a personas de distintas partes del mundo con ciertos dilemas morales (el problema del tranvía) como base para implementar ciertos tipos de moral (utilitarismo vs deontologismo) en sistemas artificiales. Otros aplican viñetas o escenarios con efectos marco para ver si el grado de atribución de culpa a humanos y máquinas es el mismo o si atribuimos la capacidad de tomar decisiones morales a sistemas artificiales (Malle, Magar y Scheutz 2019). Parece claro dadas las evidencias de la ciencia cognitiva (Bigman et al. 2019) que solo atribuiremos responsabilidad y agencia moral a robots cuando percibamos que tienen suficiente autonomía para actuar sin input humano y percibamos que tienen mente plena.

En principio, no hay ninguna razón a priori para que no se pueda crear moralidad artificial. Como tampoco hay ninguna razón lógica que contradiga la posibilidad de que se genere consciencia en un substrato físico basado en el silicio en lugar de un substrato físico basado en el carbono. De hecho la evolución por selección natural ha sido capaz de crear máquinas conscientes: nosotros. Y lo mismo se podría decir de estados afectivos (aunque reproducir procesos bioquímicos que dan lugar a los estados afectivos en sistemas electromecánicos parece imposible). Sin embargo, como veíamos más arriba en la tabla, hasta donde yo sé, no hay ninguna línea de trabajo por parte de grupos de investigación, investigadores o autores que intenten crear consciencia en máquinas o estados afectivos y haya réditos de ello por nimios que sean. Sí que es cierto que en el campo de la robótica social, distintos investigadores buscan que los robots y máquinas sepan reconocer expresiones faciales de la emoción (Breazeal 2002, Picard 1997) para permitir una interacción acorde con la humana y también hay polémicas en el debate teórico-filosófico sobre la consciencia en máquinas (Gamez 2008). Pero una cosa es equipar máquinas con sensores que monitoricen rasgos y señales humanas y respondan apropiadamente a estas señales y/o teorizar sobre la posibilidad de consciencia en máquinas; y otra cosa bien distinta es crear en máquinas fenomenología, qualia, estados afectivos o experiencias subjetivas *sensu stricto*.

Sin embargo, no hay razón alguna que me haga pensar que la moralidad no sea computable, aunque sí que hay argumentos para pensar que la consciencia no es algorítmica (Penrose 1995) y puede que sin consciencia no sea posible la moral. Vista la comparativa de similitudes y diferencias entre la moral artificial y la moral humana creo que la ética para máquinas tiene todavía un largo camino que recorrer. En primer lugar, debe identificar y clasificar todas las potenciales competencias que integran la multifacética capacidad de la moral humana para poder diseñar una arquitectura computacional que establezca el modo en que se relacionan entre sí estas competencias y así poder reproducir la moral artificialmente. En segundo lugar, debe elegir qué enfoque metodológico quiere seguir, que sea lo suficientemente comprehensivo para conseguir el objetivo que se plantea. Y, finalmente, no debe abandonar el diálogo con la filosofía y los expertos en ética para reconocer las dificultades epistémicas y técnicas del programa general de la ética para máquinas, que no es otro que crear agentes morales artificiales.

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

Bibliografía

Abney, K. (2014): "Robotics, ethical theory, and metaethics: A guide for the perplexed" En P. Lin, K. Abney, & G. Bekey (Eds.), Robot ethics: The ethical and social implications of robotics (pp. 35–52). Cambridge, MA: MIT Press.

Alexander R. (1984/2017): *The Biology of Moral Systems*. New York. Routledge

Anderson M., Anderson S. y Armen C. (2006): "An approach to computing ethics". *IEEE Intelligent Systems*. 21 ,4, 56-63.

Angwin J. (2015): *Dragnet Nation: A Quest for Privacy, Security, and Freedom in a World of Relentless Surveillance*. New York. St. Martin Press.

Anscombe G (1957): *Intention*. Oxford. Basil Blackwell.

Arpaly N. (2004): *Unprincipled Virtue: An Inquiry Into Moral Agency*. New York. Oxford University Press.

Artal P. et al. (2017): "Visual acuity in two-photon infrared vision". *Optica*, 4, 12, 1488-1491.

Asaro, P. (2016): "The liability problem for autonomous artificial agents, in ethical and moral Considerations in non-human agents", 2016 AAAI Spring Symposium Series

Asimov I. (1950): "Runaround". En *I, Robot* (The Isaac Asimov Collection ed.). New York City. Doubleday.

Awad E. et al. (2018): "The moral machine experiment". *Nature*. 563, 59-64

Bard N. et al. (2019): "The Hanabi Challenge: A new frontier for AI research". *Artificial Intelligence* X.XX-XX

Baron Cohen S. et al. (1985): "Does the autistic child have a "theory of mind"?" *Cognition*. 21, 1, 37-46

Bentham J. (1789/1996): *An Introduction to the Principles of Morals and Legislation*. Oxford. Oxford University Press.

Bigman Y. et al. (2019), "Holding robots responsible: The elements of machine morality". *Trends in Cognitive Sciences* doi:10.1016/j.tics.2019.02.008

Breazeal C. (2002): *Designing Social Robots*. Cam. Ma. MIT Press

Brembs B (2010): "Towards a scientific concept of free will as a biological trait: spontaneous actions and decision-making in invertebrates". *Proceedings of the Royal Society B. Biological Sciences* 1-10.

Bringsjord, S., y Taylor, J. (2014). "The divine-command approach to robot ethics". En P. Lin, K. Abney, & G. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 85–108). Cambridge, MA: MIT Press.

Bryson J. (2018): "Patiency is not a virtue: the design of intelligent systems and systems of ethics2. *Ethics and Information Technology*. 20, 1, 15–26

Byrne R. (2017): "Counterfactual thinking: From logic to morality". *Current Directions in Psychological Science*. 26, 4, 314–322

Caruso G. (2015): "Free Will Eliminativism: Reference, Error, and Phenomenology" *Philosophical Studies* 172 (10), 2823-2833.

Christakis N. (2019): "How AI will rewire us" *The Atlantic* [Último acceso 05/03/2019] https://www.theatlantic.com/magazine/archive/2019/04/robots-human-relationships/583204/

Christensen-Dalsgaard, J. y Carr, C. E. (2008): "Evolution of a sensory novelty: tympanic ears and the associated neural processing". *Brain Res. Bull*. 75, 365-370

Churchland P. (2017): "The Brains Behind Morality" *Cerebrum* http://dana.org/Cerebrum/2017/The_First_Neuroethics_Meeting_Then_and_Now/

Cosmides, L. & Tooby, J. (2004): "Knowing thyself: The evolutionary psychology of moral reasoning and moral sentiments". En R. E. Freeman and P. Werhane (Eds.), *Business, Science, and Ethics*. The Ruffin Series No. 4. (pp. 91-127). Charlottesville, VA: Society for Business Ethics.

Costa A. et al. (2014): "Your morals depend on language". *PLOSOne* https://doi.org/10.1371/journal.pone.0094842

Curry O., Mullins D. y Whitehouse H. (2019): "Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies" *Current Anthropology* 60,1, 47-69

Curry O. S. (2016): "Morality as Cooperation: A Problem-Centred Approach", En T.K. Shackelford y R.D. Hansen (eds), *The Evolution of Morality* p. 27-51 Springer International Publishing.

Crawford K y Whittaker M. (2016): *The AI Now Report, The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*, Washington, artificialintelligencenow.com.

Cushman F et al. (2010): "Our Multi-System Moral Psychology: Towards a Consensus View". En: Doris, J.M., (ed), *Oxford Handbook of Moral Psychology*, Oxford University Press, Oxford.

Davidson D. (1980): *Essays on Actions and Events*. Oxford. Clarendon Press

Dennett D. (1997): "When HAL kills, Who´s to blame? Computer ethics" En D. G. Stork (Ed) *Hal´s Legacy: 2001´s Computer as Dream and Reality*. Cam. Ma. MIT Press.

Dennett D. (2019): "What can we do?" En John Brockman (ed) *Possible Minds: 25 Ways of Looking at AI*. pp. 41-54. New York . Penguin.

Dennis L. y Fisher M. (2018): "Practical challenges in explicit ethical machine reasoning". En International Symposium on Artificial Intelligence and Mathematics, *ISAIM 2018*, Fort Lauderdale, Florida, USA.

Gordon J.S. (2019): "Building moral robots: Ethical pitfalls and challenges". *Science and Engineering Ethics* doi: 10.1007/s11948-019-00084-5

Greene J. et al. (2009): "Pushing moral buttons: The interaction between personal force and intention in moral judgment". *Cognition* 111, 3, 364-371.

Greene J. (2013): *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. London. Penguin.

Greene J. (2015): "The raise of moral cognition". *Cognition*, 135, 39-42

Haid J. (2001): "The emotional dog and its rational tail: A social intuitionist approach to moral judgment". *Psychological Review*. 108, 4, 814-834.

Hernandez-Orallo J. (2017): *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge. Cambridge University Press.

Kant I. (1785/2006): *Fundamentación de la Metafísica de las Costumbres*. Madrid. Tecnos.

Kitcher P. (2011): *The Ethical Project*. Cambridge. MA. Harvard University Press.

Kohlberg, L. (1969): "Stage and sequence: The cognitive development approach to socialization". En D. A. Goslin (Ed.). Handbook of socialization theory pp. 347-480. Chicago, IL, Rand McNally.

Lagnado D. y Gerstenberg T. (2017): "Causation in legal and moral reasoning" En Michael R. Waldmann (ed) *The Oxford Handbook of Causal Reasoning* pp. 565-601. Oxford. Oxford University Press.

Larson J. et al. (2016): How we analyzed the COMPAS recidivism algorithm, URL http: //www.propublica.org/article/ how-we-analyzed-the-compas-recidivism-algorithm

Latorre J. (2019): *Ética para Máquinas*. Barcelona. Ariel.

Lewis M. (2014): *Flash boys: A Wall Street revolt*. New York, WW Norton & Company.

Malle B. y Scheutz M. (2019): "Learning how to behave: Moral competence for social robots". En O. Bendel (Ed.), *Handbuch Maschinenethik* \[Handbook of machine ethics\], Springer Reference Geisteswissenschaften. Wiesbaden.

Malle B., Magar S. y Scheutz M. (2019): "AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma". En I. Aldinhas Ferreira, J. Silva Sequeira, G. S. Virk, E. E. Kadar, and O. Tokhi (Eds.), *Robots and well-being*. Cham, Switzerland: Springer Verlag.

ISSN 1989-7022

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147

Mikhail J. (2007): "Universal moral grammar: theory, evidence and the future". *Trends in Cognitive Science*. 11, 4, 143-152.

Moor, J. H. (2006): "The nature, importance, and difculty of machine ethics". *IEEE Intelligent Systems*. 21(4), 18–21.

Narvaez, D. & Mrkva, K. (2014): "Creative moral imagination". En S. Moran, D. H. Cropley & J. C. Kaufman (Eds.), *The Ethics of Creativity* (pp. 25-45). New York, NY: Palgrave MacMillan

Nucci, L. P., & Turiel, E. (1978): "Social interactions and the development of social concepts in pre-school children". *Child Development*, 49, 400-407

Shao, C. et al. (2018b): "Anatomy of an online misinformation network". *PLoS ONE*, 13(4):e0196087.

Penrose, R. (1995): *Shadows of the Mind*. London. Vintage

Piaget, J. (1932): *The Moral Judgment of the Child*. London: Kegan, Paul, Trench, Trubner & Co.

Picard R. (1997): *Affective Computing*. Cam. Ma. MIT Press.

Rabinowitz N. et al. (2018): "Machine theory of mind". *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80, 4218-4227

Singer P. W. (2009): *Wired for War: The Robotics Revolution and Conflict in the 21st Century*. London. Penguin

Singer P. (1981/2011), *The Expanding Circle: Ethics, Evolution, and Moral Progress*. New Jersey. Princeton University Press.

Tomasello M. (2018): "The normative turn in early moral development". *Human Development*. 61, 248–263

Tomasello M. (2019): *Becoming Human: A Theory of Ontogeny*. Cam Ma. Harvard University Press.

de Waal F. (2016): *Primates and Philosophers: How Morality Evolved*. New Jersey. Princeton University Press.

Vanderelst D. Winfield A. (2018): "The dark side of ethical robots". AIES '18 Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 317-322.

Wallach W., Allen C. y Smit I. (2007): "Machine morality: bottom-up and top-down approaches for modelling human moral faculties". *AI & SOCIETY*, 22(4), 565–582.

Wallach, W., Franklin, S., y Allen, C. (2010): "A conceptual and computational model of moral decision making in human and artifcial agents". *Topics in Cognitive Science*, 2(3), 454–485.

Notas

1. Hasta ahora solo existen robots "buenos" y hay muchos seres humanos "malos". Es muy probable que cuando seamos capaces de crear robots y otras manifestaciones de Inteligencia Artificial autoconscientes, tengamos que pedir nosotros (los seres humanos) disculpas. Actos vandálicos contra robots de seguridad, desmembramientos de androides, decapitaciones de autómatas… conducen a la pregunta: ¿Por qué causamos daño a los robots? Véase, estas páginas web: http://stoprobotabuse.com/ y http://www.aspcr.com/ para mayor información sobre el movimiento que hace campaña contra los abusos hacia los robots. No olvidemos que la historia humana está llena de actos entre humanos que desafían los principios que la reflexión ética considera aceptables, mientras que todavía no hay robots que hayan sido inmorales, es decir, "malos".

2. Un informe de la Unión Europea (Commission on Civil Law Rules on Robotics (2015/2103(INL)) sugiere que máquinas capaces de auto-aprendizaje deben tener un estatus legal de "personas electrónicas". La idea no es conceder DD.HH. a los robots, sino una cuestión práctica de eximir de responsabilidad por las acciones de máquinas inteligentes a los dueños de las empresas. Sin embargo, en una carta a la Comisión (https://g8fip1kplyr33r3krz5b97d1-wpengine.netdna-ssl.com/wp-content/uploads/2018/04/RoboticsOpenLetter.pdf), 156 expertos de 14 países alertaban de los riesgos y peligros de adoptar la propuesta del Parlamento Europeo. No obstante, la expansión del circulo moral (Singer 1981/2011) ha seguido creciendo y añadiendo

a los animales bajo el criterio de sintiencia. Recientemente, también a ecosistemas y plantas como los derechos conferidos a lagos y bosques. ¿Por qué no incluir en el círculo moral a robots?

3. Para no complicar el tratamiento que aquí quiero ofrecer de la noción de agencia moral, asuma el lector por un propósito expositivo la posibilidad de autonomía y libre albedrio, aunque, por supuesto, son conceptos en disputa en la literatura de investigación con posiciones que van desde la defensa de la existencia del libre albedrio (Brembs 2010) hasta su negación (Caruso 2015). Según qué posición uno mantenga sobre la metafísica del libre albedrio tendrá sentido hablar de agencia moral o no.

ISSN 1989-7022

Dilemata, año 11 (2019), nº 30, 129-147