VI CONGRESO LATINOAMERICANO DE FILOSOFÍA DE LA EDUCACIÓN
BOGOTÁ, JULIO 12 - 14 DE 2023
Hospitalidad y reencuentro: volvernos a ver para pensar el sentido de la educación y de la filosofía

Enseñanza de modelos de razonamiento práctico a agentes artificiales o sobre una ética para la inteligencia artificial

Miguel Fonseca Martínez
Universidad La Gran Colombia
miguel.fonseca@ugc.edu.co

Resumen

El impacto de los desarrollos de la Inteligencia Artificial (IA) en el desarrollo de la humanidad, en un corto y mediano plazo, suponen el planteamiento de reflexiones fundamentales sobre la relación de lo humano con este tipo de sistemas y artefactos. Las preguntas que ineluctablemente surgen son, entre otras: ¿Cómo debemos relacionarnos con dichos sistemas?; ¿cuáles son los alcances de la acción de la IA para con nosotros y nuestro entorno?; ¿Cómo deberían comportarse tales artefactos? En definitiva, dichas preguntas se sintetizan en el problema de qué debemos hacer con ellos y qué deberían hacer ellos con respecto a nosotros.

Las respuestas a estos interrogantes dependen de nuestra concepción ontológica de la IA, en tanto esta sea la de un conjunto de objetos o de sujetos. Desde la segunda concepción se derivan problemas éticos como las denominadas éticas de las máquinas y la agencia moral artificial.

En el presente trabajo se asumirá como campo de la reflexión la segunda perspectiva o concepción, y se entenderán a los sistemas y artefactos de la IA como sujetos doxásticos con capacidad relativa de razonamiento práctico. Dada esta premisa se buscará comprobar, en el marco de una IA débil, que existen modelos de razonamiento práctico susceptibles de ser enseñados a agentes doxásticos artificiales (Fonseca, 2020). Como soporte básico del argumento se expondrá la construcción de un modelo de razonamiento práctico artificial basado en la ética de la virtud de Tomás de Aquino (Aquino, 1981), como un pretexto y recurso heurístico relevante.

Palabras clave: Enseñanza, ética, inteligencia artificial, agencia moral artificial, agencia eidética, Tomás de Aquino.

Resumo

O impacto dos desenvolvimentos da Inteligência Artificial (IA) no desenvolvimento da humanidade, no curto e médio prazo, supõe a abordagem de reflexões fundamentais sobre a relação do humano com este tipo de sistemas e artefactos. As questões que inevitavelmente surgem são, entre outras: Como nos devemos relacionar com estes sistemas?; Qual o alcance da ação da IA em relação a nós e ao nosso meio ambiente?; Como esses artefatos deveriam se comportar? Em última análise, estas questões são sintetizadas no problema do que devemos fazer com eles e o que eles devem fazer em relação a nós.

As respostas a estas questões dependem da nossa concepção ontológica de IA, desde que esta seja a de um conjunto de objetos ou sujeitos. Da segunda concepção surgem problemas éticos, como a chamada ética das máquinas e a agência moral artificial.

No presente trabalho, a segunda perspectiva ou concepção será assumida como um campo de reflexão, e os sistemas e artefatos de IA serão entendidos como sujeitos doxásticos com relativa capacidade de raciocínio prático. Dada esta premissa, procuraremos verificar, no quadro de uma IA fraca, que existem modelos de raciocínio prático que podem ser ensinados a agentes doxásticos artificiais (Fonseca, 2020). Como suporte básico do argumento, será exposta a construção de um modelo de raciocínio prático artificial baseado na ética das virtudes de Tomás de Aquino (Aquino, 1981), como pretexto e recurso heurístico relevante.

Palavras-chave: Ensino, ética, inteligência artificial, agência moral artificial, agência eidética, Tomás de Aquino.

Abstract

The impact of the developments of Artificial Intelligence (AI) in the development of humanity, in the short and medium term, suppose the approach of fundamental reflections on the relationship of the human with this type of systems and artifacts. The questions that inevitably arise are, among others: How should we relate to these systems?; What are the scope of the AI action towards us and our environment?; How should such artifacts behave? Ultimately, these

questions are synthesized in the problem of what we should do with them and what they should do with respect to us.

The answers to these questions depend on our ontological conception of AI, as long as this is that of a set of objects or subjects. From the second conception, ethical problems arise, such as the so-called ethics of machines and artificial moral agency.

In the present work, the second perspective or conception will be assumed as a field of reflection, and AI systems and artifacts will be understood as doxastic subjects with a relative capacity for practical reasoning. Given this premise, we will seek to verify, within the framework of a weak AI, that there are models of practical reasoning that can be taught to artificial doxastic agents (Fonseca, 2020). As a basic support for the argument, the construction of an artificial practical reasoning model based on the virtue ethics of Thomas Aquinas (Aquino, 1981) will be exposed, as a relevant pretext and heuristic resource.

Keywords: Teaching, ethics, artificial intelligence, artificial moral agency, eidetic agency, Thomas Aquinas.

IA: Definición y Fundamentos

La IA puede ser definida como: "El campo dedicado a la construcción de animales y personas artificiales, o al menos que en ciertos trasfondos son creaturas artificiales que parecen ser animales o personas dependiendo de su desempeño agencial" (Bringsjord y Govindarajulu, 2018, p.1). En un primer momento la disciplina, o más bien el conjunto de disciplinas que la constituyen, asumió el reto de la singularidad, es decir, crear entes que simularan de forma indiscernible la inteligencia humana. El denominado *Test de Turing*, que aparece formulado en el texto pionero "Computing Machinery and Intelligence" (Turing, 1956), consiste en un contrafáctico, en el cual, la fuerza de las respuestas de un conjunto de preguntas configura un criterio 50/50 de indiscernibilidad entre un sujeto humano y otro artificial. Este criterio determinaría una concepción fuerte de la inteligencia artificial; el objetivo sería crear humanos artificiales. No obstante, tal empresa ha ido menguando, sobre todo por elementos pragmáticos y por límites epistémicos casi que infranqueables tales como la idiosincrasia de la conciencia intencional, y la diferencia entre semántica y sintaxis que ha señalado repetidas veces Searle (1983, 1995, 2004).

Por lo antedicho han surgido propuestas moderadas que pueden ser consideradas teorías débiles sobre la IA. Una versión débil de la IA busca proponer un conjunto de modelos de actitudes del entendimiento y de modelos de razonamiento como el núcleo de un desarrollo de artefactos de IA. En esta perspectiva la IA está fuertemente vinculada al diseño de modelos de razonamiento e inferencia. No se refiere tanto a una imitación del modo de razonamiento humano, sino a modelos ideales de racionalidad adecuados a sujetos específicos y agencias particulares (Hubert, 2016).

En el marco de esta apuesta de IA débil he propuesto una apuesta epistémica denominada *agencia eidética*, que puede fundamentar a su vez la propuesta aquí contenida de modelos de razonamiento práctico para sujetos doxásticos artificiales. La agencia eidética (Fonseca, 2020) propone que pensar, y en general, la mayoría de las tareas cognitivas, incluyendo las humanas, son agencias que requieren elementos materiales para su configuración. Ciertos artefactos que suelen ser denominados objetos abstractos o ideales -de ahí que se les denomine objetos eidéticos- son extensiones andamiadas de la mente que propician tareas cognitivas. Frente a visiones internalistas ingenuas, la propuesta propone que

los seres humanos y otros seres agenciamos eidéticamente, es decir, pensamos gracias a objetos ideales que se cubren con el ropaje de lenguajes artificiales que les permiten subsistir y que, además de ser puras extensiones andamiadas de un sujeto hacia el mundo, también parecen estar determinadas como extensiones andamiadas del mundo en nuestra mente.

Así, el uso de ciertas estructuras formales puede constituir fundamentos que modelen el razonamiento en cualquier ente doxástico. Apoyándose en la denominada epistemología formal que se encarga de los problemas filosóficos relativos al conocimiento utilizando herramientas formales que tienen su origen en ciencias como la lógica y la matemática (Hendricks, 2006), la agencia eidética postula construir modelos de conocimiento a partir de la construcción de artefactos formales.

Desde este marco, el propósito general de la IA consistiría entonces en desarrollar modelos conceptuales, procedimientos de reescritura formal de dichos modelos y la construcción de estrategias de programación de máquinas físicas que puedan, con eficiencia y la mayor exhaustividad, reproducir o constituir tareas cognitivas análogas a los sistemas biológicos inteligentes (Marín y Palma, 2008).

El proceso consiste, por ello, en la posibilidad de modelar, formalizar y programar. En nuestro caso, el proceso epistémico protagónico, susceptible de cumplir con las etapas anteriormente descritas, es el conjunto de procesos de aprendizaje y su uso efectivo en el ámbito del razonamiento práctico.

Razonamiento Práctico e IA

El razonamiento, en sentido amplio, es un proceso inferencial. La entrada consiste en conjuntos de actitudes del sujeto tales como creencias, afectos e intenciones. A través de ciertos mecanismos epistémicos se produce una modificación en dichas actitudes. Generalmente este proceso es conducido por estructuras paradigmáticas o modélicas. En el caso del razonamiento práctico las actitudes modificadas son usualmente intenciones (Broome, 2013). Se puede decir que un razonamiento práctico es una respuesta razonable a las intenciones, entendidas como oportunidades (posibilidades de actuar), en un tiempo, que para el humano es limitado. No obstante, en el razonamiento práctico también existe la posibilidad de modificar inferencialmente creencias y, por esto, se puede hablar de aprendizaje; en el proceso inferencial

se acomodan nuestras creencias con respecto a agendas y acciones tanto normativas como con respecto a sus referentes en acciones individuales. El proceso deliberativo del peso de nuestras razones para la acción reorganiza la estructura inferencial en la que concurre el razonamiento práctico particular (Fonseca, 2023).

Ahora bien, los agentes resuelven esto en la deliberación interna inicial y no poco compleja del qué deberían hacer, y en un segundo momento deben resolver a la inquietud de qué van a hacer en un caso particular. Algunas teorías restringen el ámbito del razonamiento práctico a una sola de estas dimensiones; tras definir qué debería ser decidido no quedaría más razonamiento práctico por cumplir (Wallace, 2020, p.1). Como puede entreverse, en esta propuesta, el razonamiento práctico se refiere a los dos momentos. Esto principalmente por su utilidad como mecanismo de aprendizaje tanto para sujetos doxásticos naturales como artificiales.

El curso de la reflexión se dirige entonces a la suposición de que todo sistema de inferencias requiere una maquinaria lógica para obtener sus objetivos (Brandom, 1994). Así, de la misma manera que los seres humanos requieren de un sistema lógico que explicita sus prácticas de inferencias morales, un sistema artificial puede igualmente recibir un sistema lógico que modele los razonamientos prácticos, y más específicamente la resolución de agencias morales tanto en el nivel del debería como en su aplicación particular a instanciaciones del modelo deóntico para la agencia.

La premisa se refiere entonces a suponer que podemos modelar y enseñar agencias cognitivas morales que reducen lenguajes naturales y prácticas implícitas a lenguajes formales; sistemas semánticos a sintácticos; conocimiento representacional a arquitecturas de mero sentido y cálculo; procesos biológicos a circuitos electicos de soporte.

El campo de la IA simbólica o representacional es entonces el instrumento más relevante para apoyar esta tarea. En tanto el lenguaje natural, desde el cual surge la ontología que constituye como formas supervinientes a las virtudes, los derechos, las obligaciones, las normas, los sistemas culturales y en sentido amplio la moral (Tomasello, 2019), está constituido por hechos y reglas, podemos entonces construir sistemas artificiales de toma de decisiones artificiales que determinen relaciones de hechos con seguimientos de reglas morales.

El objetivo apunta a la enseñanza supervisada de teorías computables del conocimiento relativo al razonamiento práctico a sujetos doxásticos artificiales. Esto deriva en el diseño y desarrollo de modelos y técnicas de representación, inferencia, aprendizaje y control predictivo para dichos sujetos artificiales.

A esta lógica se le ha denominado recientemente en IA como Ingeniería del conocimiento:

> Como alternativa a la materia y la energía, el nuevo objeto formal de la Ingeniería del Conocimiento es el conocimiento y este; como la información, es pura forma, sólo usa la energía como soporte, pero el mensaje está en la estructura relacional y en el consenso entre los distintos observadores externos que deberán de dotar del mismo significado a los símbolos formales y físicos que constituyen un cálculo (Marín y Palma, 2008, p.7).

Sin embargo, de ello brota la necesidad de una mirada filosófica para la IA, en tanto en cuanto surge la necesidad de una teoría del conocimiento. La filosofía puede asumir como una de sus tareas la construcción de modelos de razonamiento pertinentes y adecuados para construir relaciones ventajosas entre sujetos doxásticos humanos y sujetos doxásticos artificiales. Así, razonar prácticamente en el paradigma simbólico de la IA es equivalente a especificar un conjunto de reglas de manipulación de los conceptos de entrada al sistema basado en conocimiento que genera el resultado del razonamiento, a saber, la toma de decisión (Marín y Palma, 2008, p.14).

Lógica Representación y Paradigmas de Razonamiento en IA.

Los principales paradigmas con relación a las formas de razonar en el ámbito de la IA que pueden ser revisados por la filosofía práctica se pueden comprender sinópticamente como:

- Paradigma situado: posible reducción al concepto de causalidad. Función de decisión basada en la inferencia y el control.
- Paradigma conexionista: Redes de neuronas artificiales. Parametrización a través de representaciones numéricas.
- Paradigma híbrido: Sincretismo metodológico que contiene a los dos paradigmas anteriores y otras herramientas.

La propuesta que aquí se presenta se enmarca, por lo tanto, en el paradigma situado del razonamiento y puede extenderse ulteriormente a mecanismos híbridos como robots de cuidado, por ejemplo.

Representar formalmente el conocimiento para nuestra agenda consiste en definir la centralidad de las características importantes para el cumplimiento de una agenda epistémica. La inferencia que permite el paso de premisas a conclusión se modela a través de un algoritmo y deviene en una lógica útil para el razonamiento automático. Los modelos lógicos son el marco que define este proceso. Las lógicas pueden comprenderse entonces como (SBR) sistemas basados en reglas. Estos requieren:

• Base de Conocimiento: Reglas y mecanismos de inferencias.

• Base de Hechos: Memoria de trabajo.

De ello se deriva la necesidad de una definición de regla:

Definición: Una regla es una condición para una acción.

Esta requiere:

• *Modelo conceptual: Representación de conocimientos usando estructuras no computables.*

• *Modelo Formal: Representación semi computable de los conocimientos.*

• *Modelo computable: Hace que el modelo formal sea operativo. (Base de conocimientos, motor de inferencias y estrategias de control).*

Así, podemos usar paradigmas monotónicos tan sencillos como el silogismo, basado en la deducción natural, pasando por modelos de probabilidad, redes bayesianas y modelos de causalidad basados en redes neuronales. Para nuestra heurística, usaremos el razonamiento monotónico propio de los silogismos prácticos que usó Aristóteles y Tomás para el diseño de sus sistemas morales.

El elemento creativo, por lo tanto, consiste en la decisión de determinar los formalizamos de representación dada una agenda epistémica determinada en este caso tomar decisiones virtuosas.

Razonamiento Moral para la IA

El estudio general de la IA en el ámbito del razonamiento moral se refiere a elementos deontológicos claros:

1. Qué deberíamos hacer con este tipo de sistemas.
2. Qué deberían hacer ese tipo de sistemas.

3. Cuáles son los riesgos y el control de las agencias de este tipo de sistemas u objetos.

Las preocupaciones más generalizadas con respecto a la interacción de los humanos con este tipo de sistemas se refieren a las posibles consecuencias sociales, políticas y cotidianas de algoritmos y sistemas autónomos que se comportan como cajas negras sin ningún tipo de control sobre su acción en función ni de deberes, fines, carácter o consecuencias (Harari, 2016, p.462).

Si bien la toma de decisión moral requiere como condición necesaria los mecanismos epistémicos nombrados en el acápite anterior, con respecto al razonamiento práctico, no son condiciones necesarias y suficientes; la libertad, la conciencia intencional, la deliberación interna del peso de las razones para la acción, la acracia, y todos los demás elementos que devienen de un organismo biológico, vivo, situado en un ambiente, autónomo, adaptativo y autopoiético, por el momento, no pueden ser simulados por ningún sistema artificial. Siguiendo a Tomasello (2019) la historia natural de la moralidad humana es un proceso de tal complejidad, que pretender alcanzar la singularidad de ello en máquinas es aún una utopía. La dimensión moral habita ontológicamente un tipo de ser idiosincrático y de allí igualmente es idiosincrático este tipo de comportamiento propiamente humano.

Sin embargo, análogamente, se pueden imitar ciertas condiciones que pueden enseñarse, con diferencia de grado, a sujetos doxásticos artificiales, como herramienta para una mejor interacción del hombre con estos sistemas o herramientas. Como se decía, la posibilidad de andamiar, objetivar o supervenir aspectos de la mente en el lenguaje y, sobre todo, en el lenguaje artificial, propicia la posibilidad de experimentar la enseñanza de mecanismos de razonamiento práctico supervisado a sujetos doxásticos artificiales.

Como se nota, la presente apuesta se dirige a intentar solucionar la pregunta: ¿Qué deberían hacer ese tipo de sistemas? Más allá de estudiar el comportamiento humano en el ámbito de la creación de los sistemas de IA y los problemas de sesgo que allí acaecen; más allá de referirse a los problemas de acumulación de datos y el uso de la información para manipular comunidades y minar su decisión libre (Müller, 2021, p.2); más allá de la confianza en sistemas que, en el caso del aprendizaje autónomo se comportan como cajas negras sin control; más allá del posible reemplazo de la mano de obra por este tipo de máquinas; se trata de construir técnicas de enseñanza de modelos de control moral andamiado para este tipo de agentes.

Elementos como el respeto a la dignidad humana, la vida, la justicia etc., pueden ser atribuidos vicariamente a este tipo de entes, para que puedan construir sistemas morales basados en dichos valores, que puedan automatizar de alguna manera el debería ser, y operar su toma de decisiones particulares en relación con los humanos, de una forma que sea ventajosa éticamente para los últimos.

A esta dimensión de la ética se le denomina *Machine ethics* (Anderson, 2007, p. 15). La ética para máquinas se funda en su comprensión de ellas como sujetos más allá de su uso como objetos por parte de los humanos. Más allá de la derrotablidad antes expuesta, se considera que se puede asegurar la enseñanza de sistemas morales que aseguren un comportamiento regulado de las máquinas con relación a los sistemas morales humanos con las que se relacionan. La idea que se desprende es entonces que, con diferencias de grado, se pueden considerar agentes morales autónomos (Van Wynsberghe y Robbins, 2019).

Agentes Morales Autónomos

El compromiso y problema que se desprende de considerar a las máquinas como sujetos es la atribución de responsabilidades y derechos. El grado de desarrollo moral deviene de la posibilidad de justificación *post hoc* de su comportamiento, en el mismo sentido que lo hacen los seres humanos. Aquí, como se ha dicho, se tendrá una aproximación inicial a sujetos doxásticos supervisados y se dejará de lado el problema de la conciencia como elemento fundamental para la justificación *post hoc* de las decisiones morales.

En este sentido restringido, la justificación y responsabilidad seguirán siendo justificadas vicariamente y devendrán de los mecanismos iniciales de programación y de supervisión del aprendizaje que busca un óptimo de comportamiento en agendas diseñadas por humanos, para humanos.

La responsabilidad entonces en una responsabilidad distribuida, debido a que, de hecho, es una agencia compartida e interrelacionada con humanos (Taddeo y Floridi, 2018, p. 751).

Un modelo Heurístico: Tomás de Aquino, virtud e IA

La concepción de la virtud como un hábito operativo con un vínculo esencial con la actividad racional es la premisa mayor para la elaboración del modelo siguiendo la filosofía moral del aquinate (Aquino, 2010, p.65).

Los componentes principales de un modelo moral basado en el concepto de virtud son las virtudes, el conjunto de reglas y el conjunto de consecuencias. Enfatiza, no obstante, en la formación del carácter moral del sujeto doxástico gracias al aprendizaje y ejercicio continuo del conjunto de operaciones de toma de decisión.

La virtud puede ser interpretada como un modelo epistémico para responder agencialmente de la mejor manera posible con respecto a un objetivo. Un acto virtuoso será, por lo tanto, dar en el blanco agencial que cada virtud diseña como agencia. Cabe decir, que dar en el blanco agencial requiere reconocer las condiciones *caeteris paribus* donde puede obtenerse el mejor mundo posible para la toma de determinada decisión.

La virtud, siguiendo el espíritu tomista, es una cuestión de grado; la mejor acción posible puede ser metrizada en ránquines de bondad (Spohn, 2012). La acción ética va de lo supererogatorio al vicio como razonamiento insuficiente.

El campo de operación es el tiempo. En el caso de los humanos la virtud se ejecuta en un campo temporal finito y con una alta incertidumbre (Finnis, 2021, p.1). De esto se desprende la dificultad de acotar la relevancia de una decisión. No obstante, en el caso de una máquina esto es un asunto de menor relevancia y quizá solamente dependa de la obsolescencia del mecanismo y, sobre todo, de la relación calculada con un humano.

Ahora bien, ¿cuáles serían los elementos del comportamiento virtuoso tomista en clave de IA? Jhon Finnis aclara el particular cuando afirma:

> Moral and political philosophy for Aquinas, then, is (1) the set or sets of concepts and propositions which, as principles and precepts of action, pick out the kinds of conduct (that is, chosen action) that are truly intelligent and reasonable for human individuals and political communities, together with (2) the arguments necessary to justify those concepts and propositions in the face of doubts, or at least to defend them against objections. It is a fundamentally practical philosophy of principles which direct us towards human fulfillment (flourishing) so far as that happier state of affairs is both constituted and achievable by way of the actions (chosen conduct) that both manifest and build up the excellences of character traditionally called virtues. If one must use a post-Kantian jargon, it is both "teleological" and "deontic", and not more the one than the other (Finnis, 2021, p.1).

Nuestra base de conocimiento es el conjunto de conceptos y principios necesarios para la acción. En segundo lugar, nuestro mecanismo de inferencia consiste, en este caso, en el seguimiento del silogismo práctico (basado en el *modus ponens* del SDN) que determina el actuar razonable. Se suma a este mecanismo de inferencia la posibilidad de metrizar los grados de bondad asignados, a través de un mecanismo de lógica difusa como puede ser la teoría del ranquin de Spohn (2012) que metriza los grados de confianza en una proposición desde lo supererogatorio hasta lo insuficiente.

"Definition 6.2. Let k,t, A, and B as in 6.1. Then A is a

Supererogatory		$\tau (B\|A) > \tau (B\| \neg A) > 0$
Sufficient	Reason for B w.r.t. κ iff	$\tau (B\|A) > 0 \geq \tau (B\| \neg A)$
Necessary		$\tau (B\|A) \geq 0 > \tau (B\| \neg A)$
Insufficient		$0 > \tau (B\|A) > \tau (B\| \neg A)$

" (Spohn, 2012, p.107).

Además, la teoría de Spohn, nos permite estructurar un mecanismo de aprendizaje y control del mecanismo de inferencia en tanto en cuanto permite la revisión de sistemas de creencia dada información nueva en condiciones normales.

Las bases de hechos dependerán de la agenda moral que se quiera andamiar y supervisar, o bien por bases de datos a priori, o por mecanismos perceptivos que alimenten los grados de creencia en las proposiciones morales que determinan la base de conocimiento.

Pensemos por un momento en un ejemplo o instanciación muy sencilla de este sistema moral artificial. Pensemos en un automóvil con un sistema de moral virtuoso.

Modelo conceptual (algoritmo):

Premisa Mayor 1. Todo sujeto debería conducir sobrio (virtud: cuidado de sí, prudencia, etc.).

Premisa menor 2. Al sujeto x se le pueden acusar los predicados metrizados como razones supererogatorias, suficientes, necesarias o insuficientes para afirmar que esta + o- sobrio.

Conclusión: El sistema artificial (automóvil) (no) enciende.

El sistema artificial reporta su condición al usuario y a un ente supervisor de soporte. El sistema artificial recopila la información relevante para nuevos casos posibles con autorización del supervisor de soporte (formación del carácter moral del mecanismo artificial). Metodológicamente se requiere la formación de dos tipos de modelos, a saber:

Modelo Formal: Formalizar en lógica de primer orden y teoría el ranquin.

Modelo computacional: Determinar Llave de proximidad, inicio de mecanismos de reconocimiento facial, de presión, de reconocimiento de retina, de operación de voz. Determinar métodos de correspondencia con el modelo de ranquin (probabilidad e inferencia); determinar estructuras causales para habilitar o inhabilitar encendido y ulterior mensaje de reporte al igual que la recopilación de información de aprendizaje para mejorar experiencia de reconocimiento perceptual.

El caso anterior ilustra claramente el objetivo de la presente propuesta. Quien aprende el valor del cuidado, la templanza y la prudencia es el automóvil, no quien lo usa. Si bien lo hace de manera vicaria, apoya el proceso de relación moral con el ser humano y reduce la probabilidad de comportamientos que puedan causar daños.

Cada operación, bajo este marco es modélica, y depende de los objetivos agenciales que la máquina y el humano compartan. Esto será diferente, y, a su vez la aplicación del modelo, dependiendo, por ejemplo, si es un robot de cuidado, una laptop que cuida el bienestar laboral de su usuario y muchos otros mecanismos con los que interactuamos cotidianamente. También, por ejemplo, si lo que queremos programar ahora no tiene que ver con el cuidado, sino, por ejemplo, con la justicia.

Como se decía arriba, en esta propuesta se suspenden todas las inquietudes de IA fuerte y, para el caso de la moral del Aquinate, la necesidad esencial de la libertad como condición básica del razonamiento moral. No obstante, retomando los mismos pensamientos del Doctor Angélico, no en todas las instanciaciones se realiza una decisión en sentido pleno; la justificación de la acción moral suele ser post hoc y explicativa; la mayoría de las veces operamos automáticamente debido a la estructura agencial que hemos formado como mecanismo de racionalidad práctica, nada muy diferente de lo que acabamos de proponer para las máquinas.

En lo cual no diferimos una palabra es que, tanto en un dominio como en otro, el lecho del río consiste en saber que: “El bien tiene que ser hecho y perseguido y el mal evadido” (Aquino, 1981, ST I-II q. 94 a. 2).

Bibliografía

Anderson, M. y Anderson, S. 2007. *Machine Ethics: Creating an Ethical Intelligent Agent*. AI Magazine, 28(4): 15–26.

Aquinas, T. 1981. *Summa Theologiae (A Treatise on Theology)*, Parts I [1265–8], I–II [1271–2], II–II [c.1271], III [1272–3], London: Fathers of the English Dominican Province. Cristian Classics Edition.

Aquino, T. 2010. *Comentario a la ética a Nicómaco de Aristóteles*, Pamplona: Eunsa.

Brandom, R. 1994. *Making it Explicit: Reasoning Representing and Discursive Commitments*, Cambridge: Harvard University Press.

Bringsjord, S. y Sundar, N.G. 2018. *Artificial Intelligence*. The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2018/entries/artificial-intelligence/>.

Broome, J. 2013. *Rationality through Reasoning*, Oxford: Wiley-Blackwell.

Finnis, J. 2021. *Aquinas' Moral, Political, and Legal Philosophy*. The Stanford Encyclopedia of Philosophy (Spring 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2021/entries/aquinas-moral-political/>.

Fonseca, M. 2020. *Agencia Eidética: agencia material, artefáctos y agentes eidéticos*. En: Moreno J, et al, (2020). Tecnología, Agencia y Transhumanismo. pp 39-46, Bogotá: USTA.

Fonseca, M. 2023. *Belief and Society*, Bogotá: Universidad La Gran Colombia.

Harari, Y. 2016. *Homo Deus: A Brief History of Tomorrow*, New York: Harper.

Hendricks, V.F. y Pritchard, D. (eds.). 2006. *New Waves in Epistemology*, Aldershot: Ashgate.

Huber, F. 2016. *Means-End Philosophy*, in: Freitag, W, Rott, H, Sturm H, Zinke, A. (2016), Von Rang und Namen. Philosophical Essays in Honour of Wolfgang Spohn (pp 173-198), Münster: Mentis Verlag.

Müller, V. (2021). "Ethics of Artificial Intelligence and Robotics", The Stanford Encyclopedia of Philosophy (Summer 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/

Marín, R y Palma, J. 2008. *Inteligencia Artificial*, Bogotá: Mc Graw Hill.

Searle, J. 1983. Intentionality: An Essay in the Philosophy of Mind. Cambridge University Press.

Searle, J. 2004. *Mind: A brief Introduction*, Oxford: Oxford University Press.

Searle, J. 1995. *The Construction of Social Reality*, Oxford: Oxford University Press.

Spohn, W. 2012. *The Laws of Belief: Ranking Theory and Its Philosophical Applications*, Oxford: Oxford University Press.

Taddeo, Mariarosaria y Floridi. 2018. *How AI Can Be a Force for Good*, Science, 361(6404): 751–752. doi:10.1126/science.aat5991

Tomasello, M. 2019. *Una historia Natural de la Moralidad Humana*. Bogotá: Ediciones Uniandes.

Turing, A. 1950. *Computing Machinery and Intelligence*, Mind, LIX: 433–460.

Van Wynsberghe, A. y Robbins, S. 2019. *Critiquing the Reasons for Making Artificial Moral Agents*, Science and Engineering Ethics, 25(3): 719–735. doi:10.1007/s11948-018-0030-8

Wallace, R. 2020. *Practical Reason*, The Stanford Encyclopedia of Philosophy (Spring 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2020/entries/practical-reason/>.