Moralidad Artificial

Mariana Florencia Olezza
marianaolezza@gmail.com
Rec. 10/05/2018 Apr. 05/09/2018

Resumen:
El objetivo de este trabajo es llevar el debate de la filosofía práctica, de la ética, al campo de la tecnología. Particularmente se analizará si es válido o no llamar morales a los Agentes Artificiales (AAs) –es decir, drones, autos que se manejan solos, (ro)bots–, debatible a causa de su falta de libre albedrío. En segundo lugar, se buscará, a través de un ejemplo práctico, establecer una manera de implementar la toma de decisiones éticas en los AAs, usando conceptos de la ética kantiana, la ética aristotélica y el utilitarismo para su construcción. Finalmente se establecerá cuáles son los límites de la responsabilidad por parte del equipo de desarrollo del producto en el diseño de estos sistemas.

Palabras clave: Agentes artificiales - Aprendizaje máquina - Ética - Kant - Aristóteles - Utilitarismo

Abstract:
The main goal of this article is to take the ethical or practical philosophy discussion to the field of technology. It will be of special interest to analyze whether it is valid or not to claim that Artificial Agents (AA) are "moral" –by AA we mean drones, self driving cars, (ro)bots– which is arguable due to the fact that they lack free will. The second objective will be to, by means of a practical example, look for a way to implement an ethical decision-making concerning the AAs, using concepts ranging from kantian ethics, aristotelian ethics to utilitarianism for its construction. Finally we will establish the boundaries of the responsibility that the product development teams have in the design of these systems.

Keywords: Artificial agents - Machine learning - Ethics - Kant - Aristotle - Utilitarianism

Introducción

El objetivo de este trabajo es llevar el debate de la filosofía práctica, de la ética, al campo de la tecnología. Particularmente se analizará si es válido o no llamar morales a los Agentes Artificiales (AAs), debatible a causa de su falta de libre albedrío. También se buscará establecer una manera de implementar la toma de decisiones éticas en los AAs con un ejemplo práctico. Se buscará también establecer cuáles son los límites de la responsabilidad por parte del equipo de desarrollo del producto en el diseño de estos sistemas.

Ejemplos de AAs pueden ser cualquier dispositivo que compute, pero en principio se pensará en los interactivos, autónomos y adaptables, como autos, trenes y aviones que se manejan solos, drones diversos, (ro)bots.

En la primera parte de este trabajo, se llevará a cabo un análisis acerca de los AAs, la responsabilidad y la libertad. Se discutirá la posibilidad de considerar a los AAs como agentes morales y operar sobre ellos un discurso prescriptivo.

En la segunda parte de este trabajo, se analizará un caso de AA con "aprendizaje máquina", que utiliza redes neuronales artificiales. Con este análisis se mostrará la forma de operar adquiriendo nuevas experiencias por parte del AA, sus etapas y la toma de decisiones éticas. Esto requerirá un breve análisis técnico (muy a nivel conceptual), junto con un análisis kantiano y utilitarista, y la aplicación en una última etapa del silogismo práctico aristotélico. Con esto se podrá determinar los límites de la responsabilidad por parte de los humanos involucrados, y sacar diversas conclusiones de interés.

Para la realización del trabajo en su totalidad se recurrirá a conceptos teóricos de la ética kantiana, del utilitarismo, y de la ética aristotélica.

I. Agentes Morales y Responsabilidad

Se comenzará este trabajo analizando la relación entre los agentes morales y la responsabilidad. Como comenta Ricardo Maliandi, si no se supone la libertad (al menos en el sentido de libre albedrío) del agente moral, no se le puede atribuir responsabilidad por sus actos. Y si no puede atribuírsele responsabilidad, ligada a la libertad, entonces ya no se los considera agentes morales:

> No habría "agentes" sino más bien autómatas, o marionetas, movidas por alguna fuerza que les es extraña y que no pueden resistir. (Maliandi, 2004, p. 119)

El problema de la libertad, de orden metafísico, tiene importancia para la ética normativa. Existen posturas deterministas (por ejemplo, la filosofía estoica, Spinoza) e indeterministas (por ejemplo: la filosofía epicúrea, sartreana). Un intento de conciliación entre posturas deterministas e indeterministas importante es la ética kantiana. Siguiendo a Kant en la tercera de las antinomias de la *Crítica de la Razón Pura:*[1] La tesis de ésta afirma

1 En las antinomias se dan dos proposiciones contradictoriamente opuestas de manera tal que puede demostrarse tanto la tesis como la antítesis. Estos son perfectamente correctos desde el punto de vista lógico formal. Las contradicciones aparecen cuando la razón trata de conocer más allá de los límites de la experiencia.

la libertad, y sin ella no tendría sentido todo juicio moral (Kant: 2007a). La antítesis, por otro lado, afirma que todo está causalmente determinado. En la ética kantiana si bien todos los actos realizados, en tanto fenómenos, están causalmente determinados, a la vez pueden ser expresión de un yo nouménico libre.

La fundamentación de las normas en la ética, por otro lado, puede ser deontológica o consecuencialista. La ética de Kant es deontológica al estar basada en principios morales básicos. En cambio, el utilitarismo, y la ética aristotélica son consecuencialistas o teleológicas, ya que miran el fin, las consecuencias para guiarse.

A continuación, se mostrará una definición, de Luciano Floridi, de agente moral, que no es ni consecuencialista ni deontológica.

> Definición (DEF): Una acción es calificable moralmente si y sólo si puede causar bien moral o mal moral. Un agente es un agente moral si y sólo si es capaz de llevar a cabo una acción moralmente calificable. (Floridi, 2004, p.199)

DEF ni afirma ni niega que la evaluación específica de la moralidad de un agente dependa ya sea de los resultados de las acciones del mismo, o de las intenciones originales, sus principios.

Supóngase un Agente Artificial (AA) como un robot. Los AAs pueden estar suficientemente informados, ser "inteligentes" y capaces de llevar a cabo acciones moralmente relevantes, independientemente de los humanos que los hayan diseñado, causando de esta forma "bien artificial" y "mal artificial". Tienen interactividad, autonomía y adaptabilidad. Pero no tiene sentido decir que son libres. Se entiende que sería ridículo querer "culpar" a un AA por su comportamiento. Para el enfoque tradicional, mencionado por Maliandi, la identidad de agente moral sin responsabilidad (como agente moral) es algo vacío, y por lo tanto saltando todas estas distinciones, se podría hablar de agentes moralmente responsables y de agentes morales como sinónimos.

La presuposición fundamental es que se debería reducir todo discurso prescriptivo a un análisis de responsabilidad. Esto es una falacia jurídica. De hecho, hay mucho lugar para el *discurso prescriptivo* que es independiente de la asignación de responsabilidades[2] y que por lo tanto, requiere de una clara identificación de los agentes morales.

Los buenos padres, por ejemplo, usualmente llevan a cabo prácticas de evaluación moral al interactuar con sus niños, incluso a una edad en la que estos últimos todavía no son agentes responsables.

2 La tensión entre la falta de responsabilidad por el mal causado, y el hecho de ser la fuente de la causa, es la definición de lo trágico de acuerdo a Floridi (menciona el caso de Edipo). Existen ejemplos en todos los lugares y tiempos de acciones llevadas a cabo bajo estados de conciencia alterados, por ejemplo el caso de Platón en Fedro al mencionar delirios en los hombres inspirados por los dioses, que bajo una mirada actual podríamos relacionar con situaciones de brotes psicóticos con una tendencia al delirio místico. En esa situación se podría generar grandes males y sin embargo no tener responsabilidad de ello. Sin embargo hay lugar para la prescripción.

> [...] Los perros de búsqueda y rescate están entrenados para seguir el rastro de gente perdida. Suelen ayudar a salvar vidas, motivo por el cual reciben elogios y recompensas tanto de sus dueños como de la gente encontrada, pero esto no es el punto importante. Emocionalmente, la gente puede estar muy agradecida con los animales, pero para los perros es un juego, y no pueden considerarse moralmente responsables por sus acciones. Al mismo tiempo, están involucrados en el juego moral como los jugadores principales, y de forma correcta los identificamos como agentes morales que pueden causar el bien o el mal. (Floridi, 2004, pp. 202-203)

De acuerdo con DEF se puede pensar al agente moral como responsable o no. Sobre los agentes morales en general se pueden tomar acciones prescriptivas, a pesar de que no hayan tenido responsabilidad en sus acciones. Bajo la visión tradicional, en el desarrollo de un AA, los ingenieros (como los programadores) son los moralmente responsables, ya que solo los humanos tienen libre albedrío. Si se analiza el proceso de desarrollo de software, el ciclo se compone de cinco etapas: análisis, diseño, implementación, pruebas y documentación. El software es construido por equipos. Las decisiones administrativas son tan importantes como las de programación. Los documentos de requerimientos y especificaciones juegan un gran rol en el código resultante. La exactitud del código depende de los responsables de testearlo, y mucho software depende de componentes externos cuya procedencia y validez pueden ser inciertas: más aún, el software en uso es el resultado de su mantenimiento a través de todo su ciclo de vida.

Floridi comenta, con respecto a la tecnología, que se puede detener la regresión en la búsqueda de un individuo responsable cuando algo malo ocurre, ya que se entiende que a veces las fuentes morales de bienes o males pueden no ser un humano o grupo de humanos. Cabe aclarar aquí, y esto es de vital importancia, que esto dependerá altamente del tipo de tecnología utilizada. No aplica para todas. Pero sí para casos como el software adaptativo, como se verá en la próxima sección. Éste, para determinadas tomas de decisiones, alcanza una autonomía suficiente como para desligarse de la responsabilidad del equipo que lo programó.

> […] el programador podría anticipar los posibles cursos de acción y proveer reglas que lleven al resultado deseado dentro del rango de circunstancias en que el AA será utilizado. Alternativamente, el programador podría construir un *sistema más abierto que recolecte información, intente predecir las consecuencias de sus acciones, y personalize una respuesta al desafío. Este sistema podría incluso tener el potencial de sorprender a sus programadores con aparentemente nuevas y creativas soluciones a los desafíos éticos*".[3] (Floridi, 2004, p. 210)

Se llega entonces a dos posiciones encontradas. De acuerdo a Floridi, una máquina autónoma moral debe ser tratada como si tuviera la misma existencia moral que un ser humano. Y generalmente la moralidad, como vimos con Maliandi, se predica sobre la base de la responsabilidad. En este trabajo se establecerán bases para la primera alternativa.

3 Itálicas añadidas

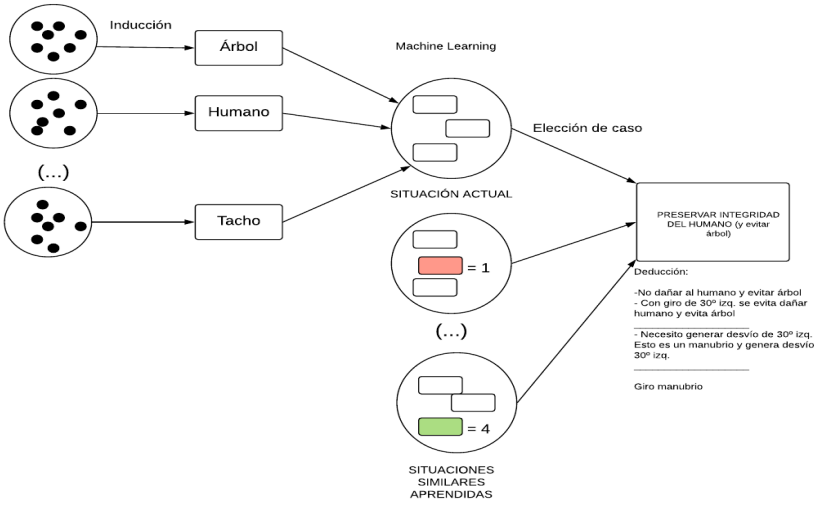
II. Implementación de Moralidad Artificial

El *machine learning* o aprendizaje máquina es un tipo de técnica de inteligencia artificial (IA) que proporciona a las computadoras (o cualquier AA) la capacidad de aprender, sin ser programadas explícitamente. Es decir, su código no es una sucesión de deducciones con cada caso posible pensado y escrito por el programador, sino que aprende a través de un proceso *inductivo*, tomando como entrada grandes cantidades de datos de muestra (de ahí que tenga una relación con Big Data).

Los algoritmos (programas informáticos) de esta manera son dinámicos y cambian en base a sus experiencias. Si el AA analizado tuviera en cambio unos conocimientos bajo los cuales actuara de manera totalmente determinada, siendo el proceder de una forma y no pudiendo ser de otra manera, se podría tener sobre él y su "accionar virtuoso" una mirada socrática ("nadie yerra a propósito"). La falla en una toma de decisiones estaría dada la información incorrecta del estado de las cosas dado por el programador. No tendría en cuenta la habituación a ciertas conductas (y tal vez "malas" conductas) por parte del AA (aprendizaje). Esto nos recuerda a la habituación de la que nos habla Aristóteles en Ética Nicomáquea.

> Y éste es el caso también de las virtudes: Pues por nuestra actuación en las transacciones con los demás hombres nos hacemos justos o injustos, y nuestra actuación en los peligros acostumbrándonos a tener miedo o coraje nos hace valientes o cobardes […] (Aristóteles, 2014, p. 161)

A continuación, se diseñará las posibles etapas de un AA con aprendizaje máquina como caso de ejemplo ilustrativo. Se supone que es un auto que se maneja solo, que se encuentra con tres posibles caminos enfrente: Uno tiene un árbol, otro tiene un ser humano, y el tercero un tacho de basura. El AA calcula que no puede frenar a tiempo y tiene que tomar la decisión de cuál de los tres caminos seguir. Se verá una posible implementación para la resolución de conflictos éticos.



Esquema

1.- Inducción

La máquina induce a través de gran cantidad de datos percibidos por el AA del ambiente, y también de sí mismo, con los que procederá a configurar los conceptos.

2.- Obtención de conceptos

Se obtienen conceptos como, por ejemplo: "árbol", "humano", "gato", "distancia", "velocidad", "cara triste", etc.

3.- Armado y análisis situacional

Es a partir de los conceptos inducidos. Esto es para seguir operando en la búsqueda de su objetivo por empezar; se recuerda al lector que el AA independientemente de su moralidad, tiene objetivos físicos que alcanzar. En esta instancia se deberá tener cargadas máximas (principios subjetivos que rigen las acciones). Las máximas de la *voluntad buena*, al decir kantiano, tienen que ser siempre *a priori*, y en este caso lo serán ya que vienen dadas por defecto, codificadas. No están en la etapa "flexible". Éstas deben cumplir con el *imperativo categórico*. Si se detecta una violación a las leyes morales simple, se puede accionar directamente en esta instancia. Si se detecta un caso más complejo como un dilema moral, se resolverá en la siguiente etapa: La propuesta kantiana, su filosofía del deber, no da una solución a los dilemas entre *deberes contrapuestos.*

Puede aplicar una máxima como "no matar", pero suponiendo que queda por resolver cuál de los otros dos caminos tomar, ya que son tres, se resolverá en la siguiente etapa.

4.- Comparación de situaciones

Si no puede seguirse simplemente una máxima, la fundamentación de las normas puede ser teleológica y de cálculo utilitarista (no se considera a los actos buenos o malos en sí mismos sino en base a sus consecuencias). Ésta es la etapa con aprendizaje máquina y redes neuronales artificiales. Se habrá entrenado previamente la red con numerosos casos de situaciones con problemas éticos, asignándole una puntuación a cada caso. Habrán sido cargadas con una 3-tupla de <situación, resolución, puntaje>. Las resoluciones más tenidas en cuenta por la red son las más similares (situaciones que apliquen) y que hayan conseguido un puntaje más favorable: "La máxima felicidad para el mayor número de personas".

En el diagrama se muestran dos ejemplos de situaciones: Una, en donde el auto impactó con un humano, con valor 1 (cuadro colorado). Y otra, donde impactó con un árbol, con valor 4 (cuadro verde). No se muestra el caso en que haya chocado con el tacho por ejemplo, donde probablemente tendría un valor más elevado y sería el elegido de no poder frenar a tiempo.

El mencionado utilitarismo es también una teoría eudemonista, al importar la felicidad o infelicidad que genera. Y en consonancia con doctrinas como el epicureísmo, asocia la felicidad al placer y a la ausencia de dolor. Al pensar en la emocionalidad, se sabe que está

ausente en los AAs. La emocionalidad tiene una relación con el actuar moralmente de tensión. Por un lado, uno debe ser sensitivo al sufrimiento de otros para actuar moralmente (empatía). Por otro lado, las emociones de los seres humanos pueden interponerse en el camino de actuar de modo moral. El problema de la *akrasia* en los AA no se presentará, ya que no sienten deseo y no se sienten arrastrados por el placer, como en el caso del incontinente o *akratés*. El problema de la empatía puede tener solución ya que existen sistemas que pueden censar los estados anímicos de las personas con la que interactúan.

Aprendizaje de ética y de lenguaje

Un ejemplo interesante de aplicación de aprendizaje máquina es el área de lenguaje y técnicas de reconocimiento del habla. Se entrena al AA hasta que por sí mismo toma decisiones respecto al lenguaje. Si se quisiera desarrollar una semántica desde cero sin "inyectar" semántica humana, sólo a través de las puras acciones del AA (como en la llamada "semántica basada en la acción" de Floridi) los AAs podrían comunicarse pero con un léxico desarrollado seguramente no humano. De ese modo la comunidad de AAs se comunicaría entre ella pero no podríamos acceder a su lenguaje. Para poder comprenderlos se tendrían que haber entrenado inyectando una semántica humana.

En este caso del aprendizaje máquina, los datos situacionales son provistos por humanos, de manera que así como los AAs heredarían características de nuestro lenguaje, en cuanto al campo de la ética heredan nuestra moralidad. Serían educados por innumerables "padres" de la sociedad, como plantea Platón en su República: Todos los que hayan aportados datos para el entrenamiento de la red en este caso. También la moral puede ser dada en la etapa (3) por los programadores en forma de máximas a criterio del equipo de desarrollo. Se deja de lado la posibilidad de una "moral superior" en el accionar de los agentes: Si el equipo programa directamente en la etapa (3) máximas, introduce su criterio. Y más que nada en la etapa (4) hay un condicionamiento en base al proceder de los aportantes de experiencias.

5.- Decisión obtenida

Si de las varias alternativas posibles filtradas, la situación fuera muy compleja y varias terminaran con la misma calificación, el *ethos* mostrando su gran complejidad, se puede considerar agregar una etapa de azar:

> [...] es más corriente admitir un "azar" no meramente gnoseológico sino también óntico. En tal sentido, predomina hoy quizá una concepción del universo más cercana a la idea del *"clinamen"* de los epicúreos (según la cual los átomos sufren, en su caída, desviaciones azarosas, dejando lugar así a hechos contingentes y al libre arbitrio)

Así como la caída de los átomos determinada y las desviaciones azarosas generan el *clinamen*, el caso de la "libertad" del agente sería una función pseudoaleatoria junto con los casos seleccionados como los mejores.

Como comenta Maliandi, en la contraposición del casuismo en el que si las normas son válidas se pueden aplicar a todo acto particular, como si sólo existiera en el AA la

etapa (3) para decidir; y situacionismo, en el que cambian las situaciones por lo que no puede haber normas válidas para todas (caso de etapa (4)) se revela la estructura conflictiva del *ethos*: La tensión permanente entre lo *universal y lo particular*. Ésta tensión juega un papel muy importante en la aplicabilidad, y en el caso del modelo desarrollado en el accionar moral del AA.

6.- Deducción

Para finalizar se puede decir que se ejecutará una serie de algoritmos para pasar a la acción que pueden verse como en forma de silogismos. Su contenido será llenado con los datos de la etapa anterior. Se supone que se parte de una premisa mayor "no dañar al humano y evitar el árbol".

Silogismo Deliberativo

El silogismo deliberativo ayuda al agente a que a través razonar halle los medios más indicados para alcanzar su fin. Una vez marcado el fin, el agente mira cómo y cuál es la mejor forma de alcanzarlo. Después de deliberar, toma una decisión, la cual surge como acto voluntario de la deliberación y es de esta decisión de donde parte la acción que busca alcanzar el fin. Por ejemplo:

P.M No dañar al humano y evitar el árbol

P.M Con un giro de 30º a la izquierda no se daña al humano y se evita el árbol

_____

C. Necesito generar un desvío de 30º a la izquierda.

Se ve que el AA tiene un fin, no dañar al humano y evitar el árbol. El agente establece que generando un desvío de 30º a la izquierda es un modo de evitar esto, de ahí concluye que necesita generar un desvío de 30º a la izquierda (recomendación).

Silogismo Práctico

Nada impidiéndolo, se concluirá la acción de giro a la izquierda del manubrio a través del silogismo práctico. Es gracias a la conexión entre estos dos estados disposicionales: deseo-necesidad (premisa mayor) y creencia (premisa menor) que se produce la *acción*.

P. M Necesito generar un desvío de 30º a la izquierda

P. m Ésto es un manubrio y genera desvío a la izquierda

_____

C. Giro el manubrio a la izquierda

Ésta última etapa es determinada por los programadores y problemas aquí claramente serían errores humanos, al igual que en la etapa (3).

Conclusión

Para concluir este trabajo, se puede decir que si bien la visión predominante acerca de los agentes morales excluye tipos de agentes que no sean responsables al no poseer libre albedrío, como los agentes artificiales (AA), existen opiniones disidentes que se atreven a hablar de la posibilidad de discurso prescriptivo aún en la falta de responsabilidad; y que en el caso de los avances de hardware y software es una temática que va a ir adquiriendo cada vez mayor importancia.

También existe una visión tradicional de que uno (o más) humanos siempre pueden encontrarse como los responsables de los AA. Se destaca que la situación dependerá en gran medida de la tecnología utilizada, pero en el caso de las técnicas como aprendizaje máquina, con la utilización de redes neuronales artificiales, por la evolución del software es a veces difícil decir que un humano tenga la responsabilidad. Más allá de esto, promover la acción normativa es perfectamente razonable incluso cuando no hay responsabilidad por parte del agente, sino sólo generación de actos morales y capacidad para este accionar moral.

Se desarrolló un esquema de AA moral, y se separó las situaciones en las que claramente el programador tendría la responsabilidad de haber una falla (como en la parte del silogismo aristotélico, donde se basa en una función con instrucciones que hacen las veces de premisas para finalizar en una acción), de las situaciones de aprendizaje máquina, donde el AA toma decisiones sobre patrones a los que nunca tuvo acceso, realizando inferencias y aprendiendo cada vez.

También se observa que de aplicar este tipo de técnicas, al enseñarle al AA en parte con experiencias de situaciones humanas (y en parte con máximas dadas como leyes), inevitablemente éste hereda la moralidad de éstos humanos en su accionar. Esto deja de lado la posibilidad de generar AAs trascendentes éticamente con respecto a los humanos. La manzana no cae lejos del árbol.

Bibliografía

— Anónimo (1967). *La vida de Lazarillo de Tormes y de sus fortunas y adversidades*. Buenos Aires: Kapelusz.

— Anderson, S.L. (2008). "Asimov's 'three laws of robotics' and machine metaethics". *AI & Soc,* 22 (477). Extraído de https://doi.org/10.1007/s00146-007-0094-5.

— Aristóteles. (2014). *Ética Nicomáquea.* Trad. Julio Pallí Bonet. Madrid: Gredos.

— García Ninet, A. (2007). "Silogismo práctico y akrasia". *A Parte Rei: revista de filosofía*, (50).

— Kant, I. (2007 [1787]) *Crítica de la Razón Pura*. Trad. Mario Caici. Buenos Aires: Editorial Colihue Clásica.

— Kant, I. (2007 [1785]) *Fundamentación de la Metafísica de las Costumbres*. Trad. Manuel García Morente. San Juan, Puerto Rico: Editorial Biblos.

— Floridi, L. (2004). *On Morality of Artificial Agents*. Cambridge: SpringerLink.

— Floridi, L. (2015). *The Philosophy of Information*. Oxford: Oxford University Press.

— Larman, C. (2004). *UML y patrones*. Prentice Hall. Lugar de edición: Monterrey, México.

— Maliandi, R. (2004). *Ética. Conceptos y problemas*. Buenos Aires: Biblos.

— Platón. (1971). *Fedro.* Madrid: Edición de Patricio de Azcárate.

— Platón. (1988). *República*. Madrid:Editorial Gredos.

— Rueda Osorio, L. *El silogismo práctico en Aristóteles*. Extraído de https://www.aca-

demia.edu/9164536/El_silogismo_pr%C3%A1ctico_en_Arist%C3%B3teles.

— Wallach W. & allen, C. (2009). *Moral Machines. Teaching robots right from wrong.* Oxford: Oxford University Press.