¿Son neutros los desarrollos tecnológicos?[1]

Are technological developments neutral?

Mg. Enrique Horacio del Carril
UFASTA; UCA; Austral

Resumen

El presente artículo revisa los criterios de atribución de responsabilidad moral por los actos dañinos de la Inteligencia Artificial. Se sostiene en él que la atribución de responsabilidad por la intención directa o indirecta del autor, que dirige la voluntad hacia un mal, no alcanza a responder adecuadamente los problemas que plantea la Inteligencia Artificial a la teoría moral.

Para ello se recurre a la diferencia entre objetos naturales y objetos culturales (artefactos) y se sostiene que, si bien hasta la irrupción del mundo digital, ésta no era una diferencia sustancial, lo es ahora por las características de las creaciones humanas (artefactos) digitales

Se propone, en cambio, que la atribución de responsabilidad podría fundamentarse en los efectos del acto de creación de la Inteligencia Artificial.

Palabras clave: inteligencia artificial; teoría moral; responsabilidad; sociedad digital; cultura.

Abstract

This article reviews the criteria of moral responsibility for the harmful acts of an Artificial Intelligence. He argues that the attribution of responsibility for the direct or indirect intention of a human author, who directs his will towards evil, does not respond to the problems that Artificial Intelligence demands from moral theory.

[1] Este artículo toma como base un trabajo del mismo título presentado en el módulo "Empleo y transformación digital" de la Diplomatura en Humanidades Digitales de la Universidad FASTA.

To argue this, the article explores the difference between natural objects and cultural objects (artifacts) and argues that, although until the development of the digital world, this was not a substantial difference, now it is.

Instead, the article proposes that the attribution of moral responsibility could be based on the effects of the act of creation of Artificial Intelligence.

Keywords: artificial intelligence; moral theory; responsibility; digital society; culture.

Introducción

Uno de los más grandes escritores de ciencia ficción escribió un volumen en el que, bajo el título "Yo, Robot" (Asimov, 2004) imaginaba una serie de situaciones en las que -en futuros imaginados- los robots dañan a las personas.

Esto no resultaría más que una de las tantas situaciones sobre robots asesinos o dictatoriales a los que la literatura y el cine nos tiene acostumbrados: HAL, Terminator y tantos otros forman parte de nuestro acervo cultural para transmitir un casi ancestral miedo a la tecnología (los mitos griegos de Prometeo y Dédalo podrían tener también esa explicación).

Lo particular de los robots creados por Asimov es que se rigen por tres reglas básicas que les impedirían dañar a los seres humanos. Estas reglas, las llamadas "tres leyes de la robótica"[2] , parecerían asegurar sin resquicios que los robots no deberían ser peligrosos… pero, precisamente, este volumen de cuentos de Asimov y otras piezas de su prolífica obra relatan situaciones en las que los seres humanos peligran porque un robot realiza determinadas acciones dañinas, pero considera que, a pesar de ello, no ha violado las tres leyes. Esto es importante de enfatizar: nunca, ninguno de los robots, se libera de las leyes o las incumple adrede.

Estas situaciones que imagina Asimov para sus robots literarios podemos trasladarlas a los adelantos tecnológicos que están ocurriendo, en toda su vertiginosidad, en nuestros días pues resultan un gran punto de partida (en general la literatura lo es) para analizar los efectos y consecuencias morales futuras (y no tanto) de la tecnología.

La neutralidad de la tecnología

La respuesta que nos surge a la pregunta sobre la neutralidad de la tecnología es, en algún punto, obvia: la tecnología no es ni buena ni mala porque los objetos, en tanto

[2] Las reglas son 1. Un robot no hará daño a un ser humano ni, por su inacción, permitirá que un ser humano sufra daño; 2. Un robot debe obedecer las órdenes dadas por los seres humanos, excepto si entran en conflicto con la primera ley, y 3. Un robot debe proteger su propia existencia en la medida en que no entre en conflicto con la primera o la segunda ley.

objetos, no son susceptibles de calificación moral. El actuar moral es inherente al actuar libre; y solo los hombres son libres. Pero, ante esta respuesta persisten, como un eco, algunas percepciones que no parecen cerrar del todo.

Algunas cosas "parecen" malas en sí: nos costaría bastante predicarle neutralidad moral a la bomba atómica, o a un virus letal creado en un laboratorio. Es cierto que, en rigor, esa maldad podría ser predicable de quien las creó y, especialmente, de quien hizo uso de ellas. Pero aun así, esta respuesta, la que se escuda en la indiferencia moral de las cosas no deja de tener un regusto amargo, insatisfactorio. Nos resulta bastante contraintuitivo considerar no-malo (y, muchos menos, bueno) a un objeto cuyo único fin visible, evidente, es causar la destrucción masiva de seres humanos.

Es que, en realidad, afirmar que las cosas no son ni buenas ni malas parece ajustarse mejor a las cosas naturales (no nos inclinamos a atribuir maldad a un terremoto o a un animal carnívoro), pero no resulta tan consistente cuando nos estamos refiriendo a los "artefactos" (Monterroza-Ríos, 2020), es decir, a los objetos culturales creados por el hombre. Y si esta respuesta no termina de cerrar respecto de muchos productos culturales, mucho menos del universo de lo digital, que parece independizarse cada vez más del hombre.

Está claro que no podemos fundar un análisis moral riguroso partiendo de cuál es nuestra percepción o nuestra inclinación espontánea respecto de un objeto o situación, pero también es evidente que las explicaciones de las teorías morales tal como las entendemos merecen un ajuste; en especial si la vinculamos con la tecnología y la proyectamos en el futuro.

Sobre el fundamento del actuar moral

El fundamento por el cual se suele afirmar que no puede predicarse de las cosas una cualificación moral es, como vimos, porque ellas no tienen una voluntad libre que pueda dirigirse hacia el bien o hacia el mal (o, con mayor precisión, hacia un bien menor).

Es cierto que un objeto puede "dirigirse" a producir un daño, pero esta trayectoria no se genera por la voluntad libre de la cosa, no elige esa opción entre las muchas posibles.

Pero esto no es exactamente así cuando nos referimos a "artefactos"; o, al menos, respecto de ellos vale intentar algunas precisiones.

Un objeto cultural, por definición, ha sido construido por el hombre; en consecuencia, en algún aspecto ha sido creado, por un lado, con un objetivo y, por el otro, por influjo de una decisión personal; y ambos son, evidentemente, actos morales. Y este objetivo y esta decisión no son indiferentes a la constitución misma del objeto.

Desde este punto de vista podría decirse que los objetos culturales reflejan, por una suerte de trasmisión existencial, la voluntad moral de su creador. Son su obra.

Esta es, creo, la razón por la que nos resistimos a predicar una indiferencia moral absoluta a un arma nuclear con potencialidad para acabar con todo el planeta. Es

cierto que ella, por sí, no es susceptible de maldad, pero el fin "natural" o "necesario" para el que ha sido concebido es, en definitiva, la destrucción de seres vivos.

Hagamos un alto a esta altura del razonamiento y retengamos esta premisa: los objetos culturales reflejan una cierta "participación" de la voluntad humana de quien los crea.

La otra observación atañe a la tecnología actual; y en especial a la Inteligencia Artificial.

Es cierto que los sesgos u otros peligros de la Inteligencia Artificial no provienen de ella misma sino de las personas que las programan. Cualquiera sea el tipo de Inteligencia Artificial que tengamos en mente (desde nuestros celulares a una ficción como Terminator), ella solo actúa en función y por causa de los datos con los que ha sido alimentada. Pero, por otro lado, no debemos soslayar un elemento importante que una Inteligencia Artificial toma decisiones autónomas (aunque no libres) a partir de esos datos con los que se alimenta. Las acciones que realiza no resultan de instrucciones de su programador, sino que son consecuencia de los datos que analiza.

Si unimos esto con la idea que acabamos de exponer, esto es, que las cosas "participan" de algún modo de la voluntad de su creador, se nos hace patente una aparente paradoja. Porque en el caso de este artefacto que es la Inteligencia Artificial, aquella "participación por transmisión" de la voluntad (buena o mala) de su creador se potencia enormemente: la Inteligencia Artificial toma decisiones por sí que, más que participar de la voluntad primigenia de quien la ha construido, la magnifican.

Pero además, es posible que estas decisiones excedan las intenciones explícitas y directas de sus creadores. Es cierto que es imposible que no estén implicadas en ese algoritmo primero, que no sean una consecuencia racional o técnica de él; pero no es una consecuencia deseada por su autor, incluso ni siquiera imaginada.

Aquí hagamos otro alto en la argumentación para anotar la segunda premisa: quienes crean una IA no necesariamente "quieren" (en el sentido de dirigir su voluntad hacia ellas) las decisiones a las que ésta llega.

Es evidente que si el artefacto se construyó para dañar, no habría una incógnita moral; o, mejor dicho, sí la habría, pero no sería más que ese problema que es tan antiguo como el mundo: el problema de la existencia del mal.

Pero éste no es el supuesto que queremos analizar aquí.

El objetivo de este trabajo no es (ni podría serlo) llegar a conclusiones definitivas, sino plantear puntos de partida para razonar problemas del desarrollo tecnológico y, en especial de la Inteligencia Artificial y su impacto en la filosofía moral. El objetivo de este trabajo no es verificar el impacto de la filosofía moral en la tecnología, sino todo lo contrario: cuál es el impacto de la tecnología en el análisis moral.

Repasemos de nuevo las premisas que nos ha dejado la disquisición precedente: los artefactos tecnológicos participan de la voluntad de su creador; los seres humanos que crean una IA (un artefacto tecnológico) pueden no querer las "decisiones" que ella tome.

Algunas conclusiones provisorias

Para encuadrar, creo, con corrección este tema, introduciremos algunas otras nociones del análisis moral.

Un objeto que produce un daño puede considerarse, desde cierta perspectiva, que ha actuado "bien". Una bala que acierta en el blanco puede juzgarse desde el punto de vista de la perfección del instrumento (el arma funciona bien) y de la acción (el tiro fue preciso) y, desde estos puntos de vista, no es relevante si el blanco es una lata de cerveza o un ser humano. Estamos aquí en el nivel de lo técnico, de la corrección o eficiencia del objeto. Éste no es, en rigor, una cualificación moral del acto (Gilson, 2000).

Es más, por el contrario, un defecto técnico o una ejecución defectuosa sí pueden tener connotaciones morales porque constituirían una infracción a la *lex artis*, al modo regular que debe ejecutarse la obra (Pieper, El concepto de pecado, 1986). Ésta es el fundamento, por ejemplo, de la ética profesional: aquel que por su situación conoce y sabe, debe actuar conforme a esas reglas.

Pero más allá de este nivel de comprensión del acto, existe una dimensión que escapa a la mera referencia "técnica", y la completa.

Es la dimensión constituida por la intención general del acto, que no se dirige a la perfección del objeto o la acción sino a su fin, que no está en la constitución misma del objeto sino en la voluntad que "lo mueve". Volviendo al ejemplo, una bomba atómica puede construirse para iniciar un genocidio o, paradójicamente, para prevenirlo[3]. Éste es el nivel de la moral.

Ese fin que crea la obra o que mueve la acción puede ser bueno o malo según si se dirige al bien de la persona (o de la humanidad) o si lo perjudica (Pieper, Las virtudes fundamentales, 1997).

Hasta aquí hemos esbozado una breve referencia a los principios generales del actuar moral en la reflexión clásica. Aceptar estas premisas, pone la cuestión en un nivel diferente al de la reflexión moral actual: no podemos explayarnos en esto, pero el relativismo difícilmente podría explicar y justificar estas diferencias (García Huidobro, 2005).

Pero lo interesante del caso es lo que pueda deducirse de la aplicación de estos principios a una Inteligencia Artificial.

Es una verdad evidente que el hombre ha creado en el devenir de su historia incontables artefactos (desde un palo aguzado para cazar al metaverso); la técnica es parte fundamental y necesaria de la propia condición humana (Ortega y Gasset, 1957).

Pero lo que diferencia de otras invenciones humanas a la digitalidad es que el hombre ha creado en ella un espacio de relación social que no tiene correlato en el mundo físico; y a la Inteligencia Artificial en que ésta tiene una voluntad propia; es

[3] Podríamos imaginar, por ejemplo, un arma de destrucción masiva construida sin intención de ser usada; por ejemplo, para amedrentar a un contrario que posee un arma igual.

cierto que es una "voluntad" técnica, aplicada a los medios y no a los fines, sin carga moral, pero no deja de ser una voluntad.

Es la primera vez, en la historia de la humanidad, que el hombre crea una voluntad autónoma.

En consecuencia (primera conclusión) podemos afirmar que la Inteligencia Artificial es ella misma la causa directa de sus operaciones y, por ello, parecería abusivo endilgarle sin más análisis las consecuencias de sus actos a seres humanos que no han querido ni previsto estos efectos, siquiera como una posibilidad[4].

Por otro lado, como la decisión autónoma de la Inteligencia Artificial no tiene un fin externo a la acción misma, es una decisión "correcta" (desde el punto de vista técnico) y, por ello, previa a una calificación moral, porque esta voluntad no es libre y, por ello, no se dirige a ese fin externo a la perfección de la obra.

Pero nuevamente esto no parece ser suficiente para que al hombre, al creador de la Inteligencia Artificial, se le atribuya o, por el contrario, se le exima de las consecuencias de los actos que aquella realice (segunda conclusión).

Antes de continuar, una breve digresión: estamos asumiendo una situación hipotética en donde los daños ocurren a pesar de que no existe un error de programación o un sesgo en los datos, sino que la Inteligencia Artificial se ha construido siguiendo con absoluta corrección las reglas de la *lex artis*. Es, en este juego hipotético, una situación en que la decisión basada en datos a la que arriba la Inteligencia Artificial es técnicamente correcta pero moralmente indeseable. Eric Sadin advierte que esta es una situación a la que nos enfrentaremos en tiempos próximos: contra la "verdad aletheica" de los datos deberemos interponer una verdad axiológica opuesta (Sadin, 2020).

Pero volvamos a la paradoja. Es evidente que en el nivel técnico la cuestión no tiene una solución porque la acción o la obra es correcta, buena; es también evidente que en el nivel moral tampoco la solución termina de cuajar porque para endilgarle los efectos de un acto o una obra a una persona, es una condición que esta lo haya querido o, al menos, imaginado como posible (Basso, 1997).

¿Cómo resolvemos esta paradoja?

Insisto, nunca el hombre había creado una voluntad que excediera la propia y que actuara de manera autónoma. Y esto obliga a repensar las variables.

Intentar constatar que un acto esté dirigido a producir un mal, que ese acto sea causado por una voluntad libre que actúa de forma directa o indirecta, por acción u omisión para que se produzca el efecto, es el método por el cual es posible atribuirle las consecuencias de ese acto a una persona determinada; pero esta vía es, en el marco de acciones realizadas por una Inteligencia Artificial, insuficiente.

El análisis debería partir de otros presupuestos.

[4] No se la han "representado" diríamos en vocabulario jurídico.

El punto central en estas reflexiones que dejo abiertas con más dudas que certezas, es que el artefacto, la inteligencia artificial conserva, en sí, algo de la voluntad creadora; y que esa condición permitiría atribuir algunos actos de la Inteligencia Artificial a un ser humano en concreto.

Conocemos esta situación de los efectos del acto creador, pero únicamente referida a Dios y sus creaturas; ese Dios que habita dentro nuestro (San Agustín) y es causa de nuestra existencia (Santo Tomás de Aquino) que no solo es la causa inicial del existir sino su causa eficiente: existimos, continuamos siendo, por influjo de ese primer acto creador. Claro que en el caso de los hombres y la Inteligencia Artificial, esta relación se verificaría de un modo análogo, imperfecto; y acorde a la naturaleza de ambos.

Esta participación de la voluntad del hombre en la Inteligencia Artificial es la ligazón necesaria para poder atribuirle los efectos del accionar de esta última.

Pero la cuestión es que éste no es ya el ámbito de la moral sino de la metafísica (Alvira, Clavell, & Melendo, 1993).

La capacidad creadora del hombre puesta en acción es la causa de la existencia de la Inteligencia Artificial, porque todo ente creado tiene necesariamente una causa. Y, en consecuencia, las acciones de la Inteligencia Artificial responden, como causa última, al acto creador de quien la “hizo ser”.

Queda abierta la cuestión de establecer cuáles de los efectos del accionar de la Inteligencia Artificial son responsabilidad de su creador.

Porque en el orden de los (otros) artefactos creados por el hombre, esta relación causal conllevaba responsabilidad (atribución moral) ya que provenía de una voluntad libre que conocía y quería directamente el efecto; el artefacto no es más que un medio, un mero instrumento.

Pero hasta ahora no nos habíamos enfrentado a la existencia de una voluntad autónoma, no humana y amoral.

El desafío es, entonces, repensar la causalidad y su relación con la responsabilidad.

El Derecho resuelve esta cuestión recurriendo a la ficción de la responsabilidad objetiva: el dueño de una “cosa” (en clave jurídica, un objeto inanimado o un animal) es responsable y debe indemnizar los daños que provoca esa cosa porque, a la vez, se beneficia patrimonialmente de él.

Esto puede solucionar problemas concretos; pero no los explica. Porque, como dice Hervada, la responsabilidad objetiva puede ser un instrumento para el ajuste por equidad, pero si se la considera un factor real de responsabilidad, implica un retorno a la barbarie (Hervada, 2008).

La clave de la pregunta no es, como se adelantó, moral sino metafísica; implica reflexionar sobre los efectos de la potencia creadora humana en el objeto creado. Nuevamente, la analogía con la creación del Hombre por parte de Dios puede servir de base para la reflexión. Dios nos creó, pero nos creó libres, por lo que no le son atribuibles las acciones contra el orden moral de sus creaturas. El hombre crea la

Inteligencia Artificial, pero no como un ente libro sino autónomo, ¿Qué grado de responsabilidad (moral, no jurídica) le cabe?

Bibliografía

Alvira, T., Clavell, L., & Melendo, T. (1993). *Metafisica*. Pamplona: EUNSA.

Asimov, I. (2004). *Yo, Robot*. (M. B. Barret, Trad.) Barcelona: Edhasa.

Basso, D. (1997). *Los principios internos de la actividad moral*. Buenos Aires: CIEB.

García Huidobro, J. (2005). *El anillo de Giges.* Santiago de Chile: Andrés Bello.

Gilson, É. (2000). *El Tomismo*. Pamplona: EUNSA.

Hervada, J. (2008). *Introducción crítica al Derecho Natural.* Buenos Aires: Ábaco.

Monterroza-Ríos, Á. D. (2020). "El papel retroalimentador de los artefactos en el desarrollo de las técnicas humanas". Trilogía (Ciencia Tecnología Sociedad).

Ortega y Gasset, J. (1957). *Meditaciones de la técnica*. Madrid: Revista de Occidente.

Pieper, J. (1986). *El concepto de pecado*. Barcelona: Herder.

Pieper, J. (1997). *Las virtudes fundamentales*. Madrid: Rialp.

Sadin, E. (2020). *La inteligencia artificial o el desafío del siglo*. Buenos Aires: Caja Negra Editora.