Ti. Futuros Comunes-REVISTA DE TECNOLOGÍAS INFORMACIONALES
AÑO IV | Nº 4 | AGOSTO DE 2024

Sobre la relevancia moral de los agentes no humanos de comunicación (ANHC)

Julián Tagnin (UNPAZ)

1. Las relaciones morales entre humanos y objetos técnicos: nuevos escenarios

Algunas preguntas que estamos dispuestos a abordar en este capítulo son: ¿en qué medida los objetos técnicos pueden ser considerados responsables de sus acciones? ¿Qué características deben tener para ser considerados agentes morales? ¿Cómo se puede conceptualizar esta agencia?

No creemos nada similar a las posturas fatalistas del determinismo tecnológico: la tecnología no tiene un poder inevitable sobre la sociedad ni determina nuestro destino de forma irreversible. Pero tampoco consideramos, como lo hace el instrumentalismo, que los objetos técnicos sean moralmente neutros.

Seguiremos el trabajo de Peter-Paul Verbeek respecto de este punto. El autor neerlandés comienza su libro *Moralizing technology* con una anécdota sobre un examen de ultrasonido que realizaron con su esposa para consultar por la salud de su hijo por nacer. Al encontrarse con una serigrafía en la sala de espera, realizada por un artista con síndrome de Down, se vieron forzados a pensar en su futuro hijo como un posible paciente. Aun cuando descartaron realizar exámenes sobre este síndrome para no juzgar tal condición como un defecto o una anormalidad indeseable, la sola existencia de la posibilidad técnica cambió la forma de su

experiencia y los marcos interpretativos que guiaron sus acciones y decisiones. Esa tecnología reorganizó las relaciones entre su hijo nonato y ellos mismos (Verbeek, 2011).

En efecto, los objetos técnicos median nuestras experiencias y, aun accidentalmente, nos enfrentan a cuestiones morales. Bruno Latour comenta que aquellas críticas que apuntan a un supuesto decaimiento moral en nuestra cultura simplemente erran en observar exclusivamente a la conducta moral humana: hay que fijarse también en los objetos. Nos rodea cada vez más la moral de los objetos. Sin embargo, este autor también considera la agencia como resultado de una red de actantes, por lo cual consideramos que desde su perspectiva los ANHC no podrían tener una agencia moral por sí mismos.

La moralización de las tecnologías se refiere al proceso de codificar valores y normas morales en los sistemas técnicos, también es el intento deliberado de diseñarlos para influir en la toma de decisiones morales humanas. Su efectividad no depende de su autopercepción, muchas veces se expresa en los sesgos implícitos de los diseñadores o ejecutores de un dispositivo tecnológico o incluso en las consecuencias imprevistas de su aplicación en un contexto social. Por ejemplo, la tecnología que nos permite prever estados de enfermedad o catástrofes naturales tiene tanta relevancia ética como cualquier pregunta respecto de cómo lidiar responsablemente con el riesgo en cualquier ámbito (Verbeek, 2011: 5). Para Verbeek, cuando la materia está moralmente cargada "diseñar es la actividad moral por excelencia [...] La ética no es más un asunto de reflexiones etéreas sino también un experimento práctico, en el cual lo subjetivo y objetivo, lo humano y lo no humano, se entrelazan" (Verbeek, 2011: 40).

Uno de los asuntos más espinosos sobre la dimensión moral de los objetos técnicos es el juicio sobre su *accountability*,[1] ¿cómo hacerles rendir cuentas por sus acciones? Ya vimos cómo Latour objetaba la posibilidad de una agencia moral por otros motivos, pero Verbeek agrega que un agente debería tener la intención de actuar de un modo específico y la libertad de realizar esta intención para ser considerado moralmente responsable. En otro trabajo, no publicado, me esforcé por traducir fenomenológicamente la intencionalidad al ámbito de los objetos técnicos, pero aún no he visto que alguien haga lo mismo con el concepto de libertad o el de conciencia. Todavía no estamos preparados científico-técnicamente para materializar estos conceptos, por lo tanto y siguiendo la premisa de avanzar desde los objetos, no podemos especular mucho más sobre las últimas condiciones que pone Verbeek.

Sin embargo, sí podemos acompañar el desarrollo tecnológico mediante un aporte al compromiso ético de diseñadores, usuarios y decisores de políticas públicas vinculadas con nuestro objeto de estudio. Este aporte será identificar puntos de aplicación de la reflexión moral para casos específicos de ANHC, que abordaremos a continuación con el fin de anticipar el impacto social de las tecnologías diseñadas. Esta anticipación es una tarea compleja, no obstante. Debido a la multiestabilidad de los objetos técnicos, su condición de tener múltiples dimensiones de interpretación que incluso pueden

1 "Accountability" no tiene una traducción literal al castellano, el término refiere a la rendición de cuentas por los resultados de una acción o decisión. Se enfoca en las consecuencias de la acción o decisión. Se suele traducir como 'responsabilidad' pero en su lengua original "responsability" se refiere más bien al deber de realizar una tarea o cumplir con una obligación, no dar cuenta por ella. Óscar Oszlak es quien señaló esta diferencia en ciencias políticas por diferencias culturales que implican diversos usos de la lengua y propuso el concepto de "respondibilidad" para evitar la ambigüedad que genera su traducción usual por responsabilidad.

superponerse, no hay relaciones unívocas entre el diseño y el rol mediador de los objetos técnicos en la sociedad. En otras palabras, no contamos con formas únicas y estables para abordarlos.

Además de la accountability, otro autor neerlandés llamado Ibo Van De Poel propone a la transparencia y la reversibilidad como metavalores a considerar obligatoriamente para mantener bajo control humano la moralidad de estos agentes (Van de Poel, 2023). El metavalor de la transparencia se refiere a la posibilidad humana de acceder, de ser necesario, a cualquier parte del proceso cognitivo, ya sea el trabajo de percepción o el de la toma de decisiones, o cualquier otra información relevante que pueda generar el agente como registro de su comportamiento. Por reversibilidad se refiere a la posibilidad de, por ejemplo, recurrir a una versión anterior, tal como hacemos con la actualización de cualquier otro programa, en caso de que la evolución del agente tenga conductas intolerables para nuestros estándares morales.

El problema principal sobre el control de los ANHC surge, para este autor, de las posibles consecuencias inintencionadas de su comportamiento. Ya sea por falta del debido cuidado en su desarrollo o empleo, por una ignorancia epistémica más profunda respecto de la previsión de su conducta o por la misma indeterminación ocasionada por factores situados fuera del control de sus diseñadores o usuarios (Van de Poel, 2023: 120). No es fácil determinar cuál de estas causas generales prima en cada caso, pero sí son categorías útiles para determinar mejor qué pudo haber pasado si resulta necesario hacer un juicio moral de la mediación de un ANHC. Como derivación de estas ideas, Van de Poel llega a la conclusión de que necesitamos extender el diseño sensible a los valores de tales sistemas a todo el ciclo vital de un agente antes que restringirlo únicamente a su diseño inicial (Umbrello y Van de Poel, 2021).

Otro aporte significativo de Van de Poel es el de tratar a las potencias cognitivas como un tipo específico de sistema sociotécnico ya que tienen autonomía, interactividad y adaptabilidad. Estas características le otorgan un grado mayor de "libertad", si intentamos empezar a adaptar ese concepto a objetos técnicos. En todo caso, los humanos estamos empezando a saber qué pueden las potencias cognitivas en el sentido en que Spinoza abría la indagación respecto de "qué puede un cuerpo".[2]

La capacidad de predecir y explicar eventos sitúa a las potencias cognitivas como agentes científicamente generativos. Esta singularidad despierta muchas preguntas sobre nuestro futuro. El conocimiento en nuestra era conduce a la paradoja, reconocida por Juan Carlos Tedesco en el ámbito pedagógico, de que cuanto más conocemos menos certidumbre tenemos (Tedesco, 2003). Y estamos ante un umbral histórico en ese sentido con los ANHC que podemos representarnos al detenernos en la siguiente sentencia: no sabemos del todo cómo saben (y probablemente sea cada vez mayor la grieta entre los dos conocimientos) ni todo lo que pueden llegar a saber las potencias cognitivas. El concepto de "saber" aquí amerita un estudio no antropocéntrico para definir con precisión a qué nos referimos por tal palabra, pero nos alcanza, por lo pronto, para plantear la incertidumbre atinente a la moralidad de los ANHC.

2 "Nadie ha determinado por ahora qué puede un cuerpo" es parte de una famosa frase del filósofo Baruch Spinoza en su tratado sobre la ética. Continúa de la siguiente manera: "(...) nadie sabe de qué forma o con qué medios mueve el alma al cuerpo ni cuántos grados de movimiento puede imprimirle y con qué rapidez puede moverlo" (Spinoza,1958: 128).

Recientes investigaciones observaron que las redes neuronales desarrollan rápidamente nuevos comportamientos cualitativos a medida que se escalan o se entrenan durante más tiempo. Tomemos a AlphaGo, la potencia cognitiva estrecha que es campeona mundial indiscutida del juego de mesa oriental Go desde marzo de 2016: ni sus desarrolladores ni los maestros del Go entienden completamente los movimientos o estrategias que el agente elige durante sus partidas.

Hay otros emergentes de conductas igualmente difíciles de comprender entre los que se cuentan las alucinaciones y el *grokking*.[3] Las alucinaciones se refieren a la generación de información que no está basada en datos reales o en el entorno circundante. Son un fenómeno relativamente común en las potencias generativas si un sistema de inteligencia artificial está entrenado con datos incompletos o poco representativos, es posible que sea más propenso a generar alucinaciones en ciertas situaciones.

Del mismo modo, si se aplican técnicas de detección de anomalías o de monitoreo continuo del rendimiento del sistema, es posible identificar patrones o comportamientos que podrían conducir a la generación de alucinaciones. Sin embargo, en muchos casos, la predicción exacta de cuándo o por qué un sistema automatizado experimentará una alucinación puede ser difícil o imposible debido a la complejidad inherente de los modelos de redes neuronales y la diversidad de factores que pueden influir en su comportamiento. La comprensión completa de los motivos y su prevención efectiva sigue siendo un área activa de investigación en el campo de estos agentes.

El *grokking* puede surgir sin regularización explícita, a partir de una anomalía de optimización que los investigadores denominan mecanismo de honda, la cual puede actuar como un regularizador implícito. En algunos casos está vinculado a cambios de fases cuando las redes neuronales desarrollan rápidamente nuevos comportamientos cualitativos a medida que se escalan o se entrenan durante más tiempo (Nanda et al., 2023). En la investigación del equipo de Neel Nanda hay un esfuerzo por intentar explicar mecánicamente estos eventos, pero la explicación se ajusta a determinado dominio de fenómenos, no a todos.

En este sentido son interesantes las reflexiones del científico computacional Mikhail Belkin, aparecidas en una excelente nota de Will Douglas Heaven para la revista del MIT:

> Hay un montón de complejidad en los transformadores. Nuestro análisis teórico está lejos de lo que pueden hacer estos modelos. (Hay muchas cosas) Que hasta hace poco nadie pensó que podían hacer. Eso significa que algo fundamental estaba faltando, hay una brecha en nuestro entendimiento del mundo (Douglas Heaven, 2024).

3 El *grokking* es un término popularizado por la novela de ciencia ficción publicada en 1961 *Extranjero en tierra extraña* (*Stranger in a Strange Land*) de Robert Heinlein. Es utilizado para referirse a una forma de comprensión profunda e intuitiva que va más allá del conocimiento intelectual. Los investigadores lo usan para explicar el salto cualitativo en la comprensión del entorno por parte de una potencia cognitiva que al término de una cierta cantidad de pasos de aprendizaje parece perdida, pero de pronto encuentra una solución acertada.

Es imposible para los seres humanos manejar racionalmente ese caudal de datos sin recurrir a estas agencias. El camino recién comienza: la teoría va detrás de su objeto. Apuesto a que llegue el momento en que los mismos agentes nos sirvan eventualmente para explicarnos en nuestros términos el propio funcionamiento de lo que entendemos como anomalías por fuera de los sobreajustes y fallas.

2. Agencia moral de los ANHC

Como señalan Verbeek y Peter Kroes, de algún modo las afirmaciones sobre la agencia moral de los artefactos tecnológicos se pretenden revolucionarias: "Después de que la Ilustración trasladara la fuente de la moralidad de Dios a los humanos, estas afirmaciones quieren moverla un paso más allá: de los humanos a las cosas materiales" (Kroes y Verbeek, 2014: 4). Esta idea tiene su correlato en la proyección de Irving John Good, el matemático y criptógrafo británico que introdujo el concepto de una "explosión de inteligencia" en su artículo de 1965 titulado "Speculations Concerning the First Ultraintelligent Machine" (Good, 1966). En él, Good habla del crecimiento exponencial de inteligencia que generaría una máquina capaz de mejorar su propia inteligencia, el progreso se aceleraría de forma explosiva, superando rápidamente la capacidad intelectual humana. Este concepto es esencialmente lo que Ray Kurzweil y otros han denominado la "singularidad tecnológica". Este proyecto ha sido justamente criticado desde sus supuestos epistémicos, y nosotros ya discutimos también los problemas asociados con ese uso del concepto de inteligencia.

Comparados con las especulaciones de Good o Kurzweil, nosotros sostenemos una teoría deflacionaria de la agencia moral de los ANHC aunque, no obstante y por eso mismo, le reconocemos agencia. Desde la "ética del acompañamiento" que propone Verbeek, debemos considerar las condiciones del momento del diseño, pero también las de la inserción social como dimensiones espaciotemporales relevantes en el análisis de la agencia moral de un ANHC.

Es notable que, como señaló Langdon Winner, los objetos técnicos fueron diseñados y programados por humanos, proceso en el cual pueden encarnar y reflejar valores humanos. Por ejemplo, una potencia cognitiva bien puede ser diseñada para priorizar la equidad, la transparencia y la privacidad. En este sentido, estos agentes pueden ser considerados portadores de valores morales y, por lo tanto, sujetos a evaluaciones morales. Esto incluye considerar cómo los sistemas de comunicación no humanos pueden perpetuar sesgos, influir en la autonomía de las personas o afectar la justicia social, incluso a pesar de los planes originales de los diseñadores. El diseño sensible a los valores (VSD, por sus siglas en inglés) propone justamente considerar y respetar los valores humanos y éticos desde el inicio del proceso de diseño hasta el momento de su aplicación. Esta perspectiva tiene una noción de los objetos técnicos como pasibles de ser juzgados como entidades morales, mas no agentes. Además, queda irresuelta la cuestión de qué pasa cuando un ANHC es creado por otro, tal como proyectaban Good o Kurzweil.

Una posición modesta de la agencia moral de los ANHC podría proponer un enfoque de responsabilidad compartida, donde tanto los diseñadores y usuarios humanos como los sistemas técnicos partici-

pan en un ecosistema moral. Se podría argumentar que los diseñadores y operadores de estos sistemas tienen una responsabilidad moral en la forma en que sus creaciones o acciones afectan a las personas.

Los ANHC (algoritmos de recomendación, las potencias cognitivas y los chatbots, entre otros) influyen directamente en nuestras decisiones y acciones comunicativas. Desde la perspectiva que adoptamos, esta influencia significa que los ANHC coconfiguran nuestras prácticas comunicativas y, por lo tanto, tienen un rol en la red de responsabilidad moral. Al reconocer que los ANHC mediatizan nuestras interacciones y decisiones, la responsabilidad moral no recae únicamente en los diseñadores o usuarios humanos, sino que se distribuye incluso entre agentes no humanos. En general, podemos decir que este cuadro se alinea con una visión deflacionaria de la agencia moral, donde no se atribuye plena agencia moral[4] a los ANHC pero se reconoce su papel significativo en las redes de influencia y se abre la pregunta por su responsabilidad.

En sintonía, la postura de Bruno Latour es que cuando los no-humanos actúan como mediadores hacen que otros actores hagan cosas. Podemos decir que él apuesta por una agencia moral robusta de ensamblajes sociotécnicos con responsabilidad distribuida. Él define a los mediadores como actores que se asocian con otros actores de tal manera que "hacen que otros hagan cosas inesperadas'" (Latour, 2005: 106). Sin embargo, considero que Latour no ofrece una gran ayuda para que podamos evaluar las implicancias de estas mediaciones por el asunto mencionado de no singularizar suficientemente a los actantes.

3. Responsabilidad de los LLMs. ¿Cómo responsabilizar a estos agentes?

Christian Illies y Anthonie Meijers sostienen una tesis moderada de la relevancia moral de los objetos técnicos y hablan de una "responsabilidad de segundo orden" (Illies y Meijers, 2009) para los mismos. Estos autores, que dialogan con Verbeek, reconocen la no neutralidad de los objetos técnicos, especialmente por su rol en el abanico de *affordances*, pero no se animan a hablar de los artefactos como agentes.[5] Pero ¿qué implica una responsabilidad de segundo orden? En el sentido que lo exponen los autores esto sería omitir la responsabilidad directa de los artefactos por sí mismo y atribuírsela no solo a los usuarios, sino principalmente a los diseñadores y creadores de los objetos técnicos. Ellos deben considerar cómo sus creaciones pueden influir en el comportamiento humano y en las dinámicas sociales, incluso si los artefactos en sí no son responsables en un sentido estricto para ellos. De hecho, esta negación de lo que Verbeek propone como una agencia compartida o

4 La plena agencia moral se refiere a la capacidad de un agente para ser considerado completamente responsable de sus acciones desde un punto de vista moral. Los podemos identificar por ciertas características que incluyen la autonomía, la racionalidad, la conciencia moral, la intencionalidad y la responsabilidad, entre otras. Como hemos visto, la autonomía de los LLMs es relativa, su racionalidad es evidente y hemos justificado su intencionalidad. La conciencia moral es un asunto más complejo porque no tradujimos la conciencia en términos no antropocéntricos, ni la empatía, por ejemplo. Sobre la responsabilidad, un aspecto clave, hablaremos más adelante.

5 Recordemos que el sentido de las *affordances* es que "los artefactos tecnológicos pueden ser moralmente relevantes, por lo tanto, al crear nuevas opciones de acciones posibles desde las cuales pueda juzgarse como más correcta una situación que otra" (Moreno, 2019).

distribuida en redes de humanos y no humanos, como nuestros ANHC, es la principal crítica que separa el pensamiento de Illies y Meijers del de Verbeek sobre el asunto (Illies y Meijers, 2014).

Desde la fenomenología sociomaterial se propone una visión integrada de la agencia, distribuida entre dos categorías, necesaria y arbitrariamente separadas en la experiencia analítica, como las de sujeto y objeto. Lucas Introna señala que lo importante en este juicio "no es centrarse en la agencia material o en la agencia humana como tal, sino más bien hacer visibles las condiciones continuas de posibilidad, la forma de estar en el mundo, que hacen posible la coconstitución de las agencias" (Introna, 2013).

Katherine Hayles señala, en el mismo sentido, que las teorías éticas suelen ser intensamente antropocéntricas e individualistas cuando en realidad un acto moral, siguiendo al utilitarismo de Jeremy Bentham,[6] debería juzgarse mediante la consideración de todas sus consecuencias y "los resultados deberían ser evaluados sistémicamente de modo que permitan reconocer que no todos los actores importantes son humanos" (Hayles, 2017: 37). Hayles les llama cognizadores (*cognizers*) no conscientes prácticamente a un conjunto definido de los mismos actores que nosotros llamamos ANHC, y sistemas técnicos no conscientes a los ensamblajes. La idea de ensamblajes cognitivos con la que trabaja la autora sugiere que la responsabilidad puede ser entendida en términos de relaciones y redes. Los ANHC, al formar parte de redes con humanos conscientes, comparten la responsabilidad derivada de esas relaciones interdependientes.

Introna cita a Karen Barad, una filósofa y física estadounidense conocida por su trabajo en filosofía de la ciencia, quien sostiene que la responsabilidad humana está inmersa en un entorno de acciones que la exceden. De alguna manera es una teoría minimalista de la responsabilidad humana, ya que exhibe los límites de los efectos de nuestras acciones voluntarias: "Somos responsables del mundo en el que vivimos no porque sea una construcción arbitraria de nuestra elección, sino porque la realidad agencial se sedimenta a partir de prácticas particulares en las que tenemos un papel en dar forma" (Barad, 2003 citada en Introna, 2014). Es decir que incidimos en un mundo ya activo, nuestro rol es importante, pero nuestra fuerza no es la única en juego. Con la emergencia de ANHC tenemos nuevos actores en este entorno que, aunque todos sean creaciones directas o indirectas de seres humanos, también tienen casos muy concretos de tomas de decisiones moralmente relevantes, como pasa con los drones y la interpretación instantánea que deben hacer para decidir atacar, dónde y cómo (Chamayou, 2016).

Como apunta el filósofo italiano Roberto Espósito, en la base de la concepción neutra y pasiva de los objetos está la tradición occidental más fuerte: la filosofía griega en el dualismo platónico; la ley romana que define a los objetos desde la propiedad y su posesión; y la concepción dualista del cristianismo que separa alma y cuerpo (Espósito, 2015). Desde su perspectiva, Espósito critica la tradicional separación ontológica que se ve reforzada por estas tres vertientes. Su pensamiento ofrece una base teórica que puede ser aplicada al desarrollo de una justicia ambiental donde humanos y no humanos son interdependientes en una comunidad más amplia que ellos mismos. Esta visión

6 Esto es, de mínima, curioso además porque el utilitarismo es metodológicamente individualista, o sea que comparte el marco más afín al antropocentrismo.

promueve una ética de coexistencia y responsabilidad compartida entre todos, además, obviamente, de bregar por el reconocimiento de la agencia de los no humanos.

Pero es necesario que pasemos de la especulación teórica hacia la reflexión pragmática. Existen legislaciones de diferentes jurisdicciones que incorporan principios de responsabilidad extendida hacia los no-humanos. Algunos ejemplos son las leyes de derechos de la naturaleza en Ecuador y Bolivia, o las políticas de sostenibilidad en la Unión Europea. Sin embargo, a nosotros nos interesan los ANHC, específicamente en este caso los LLMs. En algunos países ya existen leyes y regulaciones específicas que requieren que los chatbots se identifiquen como tales. Por ejemplo, en California, la Ley B.O.T. (Bolstering Online Transparency Act) requiere que los chatbots se identifiquen cuando interactúan con personas en línea (LA Times, 2019). Esto se hace explícitamente para que los usuarios puedan ajustar sus expectativas en consecuencia de saber con quiénes hablan. Es decir, por cuestiones de transparencia y confianza que se les pide rendir a estos agentes.

Por otro lado, el Reglamento general de protección de datos (GDPR, por sus siglas en inglés) de la Unión Europea (Unión Europea, 2016) dispone que los ANHC, como chatbots y asistentes virtuales, deben informar a los usuarios sobre la recopilación de datos, y obtener su consentimiento explícito para compilarlos en primer lugar. La Directiva sobre la privacidad y las comunicaciones electrónicas (ePrivacy Directive) complementa el GDPR y se centra en la privacidad en las comunicaciones electrónicas. Los ANHC deben asegurarse de que las comunicaciones y los datos del usuario están protegidos, implicando una responsabilidad, ya sea de primer o segundo orden, en términos de seguridad y manejo de información.

Las directrices del grupo de expertos de alto nivel en inteligencia artificial de la Comisión Europea (Comisión Europea, 2019) recomiendan que las potencias cognitivas operen bajo principios de justicia, transparencia, no discriminación y responsabilidad. ¿Cómo pedirle justicia o no discriminación a un mero objeto? La visión instrumentalista es muy limitada para entender la complejidad y la dinámica interaccional de los objetos. Nosotros creemos que las teorías que reconocen la agencia y responsabilidad de los objetos pueden conducir a sistemas más adaptables y resilientes, ya que con estas perspectivas se toman en cuenta las interacciones dinámicas y se diseñan mecanismos para manejar los efectos imprevistos y no intencionales.

Tomemos como ejemplo el comunicado de la UNESCO que recomienda principios éticos para la inteligencia artificial. Allí se habla de transparencia, explicabilidad y responsabilidad. Esto sugiere que los ANHC deben operar de manera que los usuarios comprendan sus acciones y decisiones, y que haya mecanismos para responsabilizar a los desarrolladores y operadores. ¿Pero cómo podría dar cuenta un humano del comportamiento cognitivo imprevisto de otro agente? La manera en que estos modelos llegan a sus decisiones no siempre es clara, incluso para los propios desarrolladores. Para volver a un ejemplo remanido: nadie sabe los motivos por los que AlphaGo o AphaZero deciden sus movimientos en el juego del Go. Sí, naturalmente no podrían hacer nada si cortamos la luz o rompemos sus nodos o servidores. Pero justamente atender su potencia cognitiva no es antropomorfizarlos, sino entender la propia naturaleza de los ANHC. Su rendimiento es suprahumano en determinadas

tareas e infrahumano en otras. Pero no necesitamos compararnos ni asumir que todo lo que hacen es producto de nuestra voluntad, salvo en un sentido muy general e inespecífico para su control cotidiano, sino entenderlos y regularlos según su naturaleza y nuestros intereses.

Es claro que no tendría sentido enjaular a un ANHC, por ejemplo, eso solo puede operarse con animales. Cuando hablamos de responsabilizar a un LLM, por ejemplo, en lugar de sanciones humanas tradicionales nos referimos a que la responsabilidad de los ANHC debería centrarse en la transparencia, la supervisión y la rendición de cuentas sobre las cosas que hacen, que efectivamente no son pocas y serán cada vez más en nuestro entramado sociotécnico altamente complejo. Esto implica establecer normas claras para la documentación de sus decisiones y procesos, así como implementar sistemas de monitoreo continuo para detectar y corregir errores o comportamientos imprevistos. También incluye invertir en investigación interdisciplinaria para explorar las interacciones complejas entre humanos y ANHC. Estimo que es el mejor camino para que podamos aprovechar plenamente los beneficios de estos sistemas avanzados mientras minimizamos sus riesgos y aseguramos que su integración en la sociedad sea responsable y beneficiosa para todos.

Referencias bibliográficas

Chamayou, G. (2016). *Teoría del dron*. Barcelona: Futuro Anterior.

Comisión Europea (2019). Communication: Building Trust in Human Centric Artificial Intelligence. CE, Bruselas.

Douglas Heaven, W. (2024) 6 big questions for generative AI. MIT Technology Review, 127(1), 30-38.

Espósito, R. (2015). *Persons and Things*. Cambridge: Cambridge University.

Good, I. J. (1966). Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers*, 6. Oxford.

Hayles, K. (2017). *Unthought: the Power of the Cognitive Nonconscious*. Chicago: Chicago University Press.

Illies, C. y Meijers, A. (2009). Artefacts without agency. *The Monist*, *92*(3), 420-440.

----- (2014). Artifacts, Agency, and Action Schemes. In P. Kroes y P. P. Verbeek (eds.), *The Moral Status of Technical Artifacts* (pp. 159-184). New York: Springer.

Introna, L. D. (2013). Towards a post-human intra-actional account of sociomaterial agency (and morality). In *The moral status of technical artefacts* (pp. 31-53). Dordrecht: Springer Netherlands.

Kroes, P. y Verbeek, P.-P. (2014). Introduction: The Moral Status of Technical Artefacts. En P. Kroes y P.-P. Verbeek (eds.), Philosophy of Engineering and Technology. The Moral Status of Technical Artefacts (pp. 1-9). Dordrecht: Springer. doi:10.1007/978-94-007-7914-3_1

Latour, B. (2005). Reassembling the social: An introduction to actor-network-theory. Oxford; Nueva York: Oxford University Press.

Latour, B. (2007) Nunca fuimos modernos. Ensayo de antropología simétrica. Barcelona: Siglo XXI.

Los Angeles Times (2019). California obliga a los robots a identificarse en llamadas o correos. Recuperado de https://www.latimes.com/espanol/noticas-mas/articulo/2019-07-01/efe-4013506-15569377-20190701

Moreno, J. C. (2019). Contribuciones al debate sobre la relevancia moral de los artefactos tecnológicos. *Trilogía Ciencia Tecnología Sociedad*, 11(21), 91-117, 2019. Instituto Tecnológico Metropolitano.

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. arXiv preprint arXiv:2301.05217.

Oszlak, O. (2003). *¿Responsabilización o respondibilidad?: el sujeto y el objeto de un Estado responsable*. In VIII Congreso Internacional del CLAD sobre la Reforma del Estado y de la Administración Pública. (Vol. 12). Panamá.

Spinoza, Baruch (1958). *Ética demostrada según el orden geométrico*. FCE, Buenos Aires.

Tedesco, J. C. (2003). *Los pilares de la educación del futuro*. En Debates de Educación (2003: Barcelona) [ponencia en línea]. Fundación Jaume Bofill; UOC.

Umbrello, S. & van de Poel, I. (2021). Mapping Value Sensitive Design onto AI for Social Good Principles. AI and Ethics 1 (3):283-296.

Unión Europea (2016). Reglamento General de Protección de Datos (RGPD), Reglamento (UE) 2016/679. Parlamento Europeo, Bruselas.

Van de Poel, I (2023). AI, Control and Unintended Consequences: The Need for Meta-Values. (Chapter 9) In Fritzche y Santamaría (2023), *Rethinking technology and engineering*. Springer, Suiza.

Verbeek, P. P. (2011). *Moralizing technology. Understanding and Designing the Morality of Things*. Chicago: University of Chicago Press.