ARTÍCULOS

Revista de **Filosofía**
ISSN: 0034-8244

EDICIONES COMPLUTENSE

https://dx.doi.org/10.5209/resf.81655

Hiperética artificial: crítica a la colonización algorítmica de lo moral[1]

Patrici Calvo[2]

Recibido: 24 de abril 2022 / Aceptado: 24 de septiembre de 2022

Resumen. Este estudio reflexionar pretende críticamente sobre la posibilidad de un enfoque *datificado*, hiperconectado y algoritmizado de clarificación, fundamentación y aplicación de lo moral: la *hiperética artificial*. Para ello, se mostrará la ética como un saber práctico que, preocupado por la racionalización de los comportamientos libres, ha encontrado en el diálogo entre afectados el criterio de moralidad desde el cual poder criticar tanto el conocimiento como el comportamiento. Posteriormente, se profundizará en la *etificación*, el intento de establecer procesos de transformación de la realidad social y moral en datos y metadatos computables en línea. Después, se expondrá cómo los modelos matemáticos artificialmente inteligentes están colonizando progresiva e implacablemente los procesos de racionalización con arreglo a sentido, produciendo falta de sentido, anomia y psicopatologías en las democracias maduras. Finalmente, se reflexionará críticamente sobre el diseño, aplicación y uso de algoritmos de inteligencia artificial como instrumento para establecer qué es justo y *felicitante* para una sociedad digitalmente hiperconectada.
Palabras clave: ética; inteligencia artificial; *dataficiación*; hiperconectividad digital.

[en] Artificial hyper ethics: criticism of the algorithmic colonization of morality

Abstract. This study aims to critically reflect on the possibility of a *datafied*, hyperconnected, and algorithmized approach to the clarification, foundation, and application of morality: *artificial hyperethics*. To this end, ethics will be presented as a practical knowledge concerned with the rationalization of free behaviors, which has found in dialogue among affected parties the criterion of morality from which to criticize both knowledge and behavior. Subsequently, *etification* will be explored, the attempt to establish processes for transforming social and moral reality into computable online data and metadata. Then, it will be exposed how artificially intelligent mathematical models are progressively and relentlessly colonising the processes of meaning-based rationalisation, producing meaninglessness, anomie and psychopathologies in mature democracies. Finally, a critical reflection will be made on the design, application, and use of artificial intelligence algorithms as a tool to establish what is just and *succesful* for a digitally hyperconnected society.
Keywords: ethics; Artificial Intelligence; datafication; digital hyperconnectivity.

[1] Esta publicación es parte del proyecto PID2022-139000OB-C22, financiado por MCIU/AEI/10.13039/501100011033/FEDER, UE, así como en las actividades del grupo de investigación de excelencia CIPROM/2021/072 de la Comunitat Valenciana.

[2] Universitat Jaume I
calvop@uji.es

Cómo citar: Calvo, P. (2024): "Hiperética artificial: crítica a la colonización algorítmica de lo moral", en *Revista de Filosofía* 49 (1), 71-91.

1. Introducción: la moral del termostato

En las Reith Lectures de 1984, publicadas íntegramente en *Mind, Brain, and Behaviour* (Searle, 1988), John Searle contó su desconcertante conversación con John McCarthy –quien acuñó el término Inteligencia Artificial[3] y promovió la emergencia de su disciplina científica– en una de sus conferencias. En ella, McCarthy no sólo respondió afirmativamente a la pregunta planteada por Alan Turing algunas décadas antes: ¿pueden pensar las máquinas? Además, este no dudó en aseverar con rotundidad que cualquier máquina, incluso una tan simple como su termostato, era capaz de apropiarse de "creencias propias". Ante las osadas afirmaciones de McCarthy, durante la ronda de preguntas Searle le preguntó con cierto sarcasmo cuáles eras las *creencias* de su termostato, a lo que este le respondió que su "termostato tiene tres creencias: hace demasiado calor aquí; hace demasiado frío aquí; y aquí hay una buena temperatura" (Searle, 1988: 34-35).

Más allá de lo anecdótico y simpático del caso narrado por Searle, las palabras de McCarthy permiten vislumbrar cómo y hasta qué punto la disciplina de la inteligencia artificial ha establecido desde sus inicios un proceso de apropiación y asimilación de terminología extratecnológica –biológica, social y moral– con el propósito de recrear un imaginario simbólico capaz de ungir –*tokenizar*– las máquinas dotadas de modelos matemáticos inteligentes de una suerte de naturalización y humanización, o incluso de suprahumanización. Desde la reconceptualización del término inteligencia para definir tanto la capacidad computacional de las máquinas como la propia disciplina científica, hasta la redefinición y vinculación de vocablos tan ligados al ser humano como creencia, aprendizaje, interpretación, agencia, autonomía, libertad, mente, conciencia, pensamiento, juicio, valor, justicia, redes neuronales o gobernanza ética para definir la competencia comportamental de las máquinas, se ha ido conformando ese imaginario donde los modelos matemáticos artificialmente inteligentes se equiparan a las personas o, incluso, las superan tanto física como moralmente. Esta recreación simbólica ha convertido a las máquinas en *inevitables* y *necesarias* para la supervivencia y desarrollo de todos los sistemas de seres vivos y de valores humanos: ante las limitaciones de los seres humanos para objetivar sus razonamientos y salvar el sesgo emocional en los procesos de toma de decisiones, sólo éstas están capacitadas para orientar y dirigir el rumbo de la humanidad de manera justa y *felicitante* (Matsumoto, 2018).

Al respecto, el último episodio en este largo proceso de construcción simbólica alrededor de las máquinas se halla en el actual interés de diferentes disciplinas tecnocientíficas por atribuir a la inteligencia artificial altas competencias y capacidades para analizar, interpretar y objetivar lo moral. Es decir, el actual desarrollo exponencial de los modelos matemáticos artificialmente inteligente y sus logros alcanzados en distintos campos de actividad, como la democracia, la economía

[3] Por IA como término se entiende la emulación de las capacidades humanas (también cognitivas) a través de modelos matemáticos computacionales. Por IA como disciplina científico-técnica se entiende como aquel saber que se ocupa de emular la conducta humana inteligente mediante procesos computacionales (Schalkoff, 1990).

o la salud, está alimentando la falsa creencia de que es posible y deseable sustituir la razón humana – subjetiva y sesgada– por la razón algorítmica –objetiva y neutral–[4] para un mejor discernimiento de aquello que es justo y *felicitante* para la sociedad. La principal muestra de ello, es la actual tendencia hacia el diseño y aplicación de un enfoque ético de definición, fundamentación y aplicación de saber moral basado en la datificación, hiperconectivización y algoritmización de la praxis humana en todas sus dimensiones: la *hiperética artificial*[5].

El objetivo de este estudio es reflexionar críticamente sobre el posible advenimiento de un enfoque *hiper-artifical* de pensar lo moral. Para ello, en primer lugar, se describirán las principales características de la disciplina ética y su transformación dialógica a partir de la segunda mitad del siglo XX. En segundo lugar, se profundizará en el proceso de transformación de la realidad social y moral en datos[6] y metadatos computables en línea. En tercer lugar, se expondrán críticamente las principales características y los principales impactos del diseño y aplicación de un enfoque *hiperético* de clarificación, fundamentación y aplicación de lo moral con el objetivo de desvelar el discurso falaz, interesado y altamente corrosivo que le subyace. Finalmente, en cuarto lugar, se profundizará en los principales desafíos éticos que subyacen a la *hiperética artificial* para las sociedades digitalmente hiperconectadas.

2. De la prudencia al diálogo: la transformación discursiva de lo moral

Partiendo de las ideas provenientes de los primeros escritos filosóficos, pero enriquecida con las aportaciones de otros pensadores y pensadoras a lo largo de un proceso histórico y progrediente de más de dos mil años, la ética podría definirse como aquel saber práctico que se adentra en el ámbito de las relaciones libres con el objetivo de aportar claridad conceptual, dar razón de la validez de los principios, normas, acciones y decisiones en juego y aplicar todo el bagaje alcanzado sobre las distintas esferas funcionales y relacionales para ayudar a resolver la conflictividad subyacente, orientar su desarrollo en un sentido justo y *felicitante*, mejorar su credibilidad y confianza y aumentar sus impactos positivos sobre la sociedad con la finalidad de, entre otras cosas, minimizar su vulnerabilidad y posibilitar sus proyectos de vida buena (Aramayo, 1986, 1991; Conill, 2006; Cortina, 1986, 1990; Cortina y Martínez, 1996; Cortina, García Marzá, & Conill, 2003; García-Marzá, 1992).

---

[4] Se entiende por "razón algorítmica" el discernimiento basado en "datos objetivos" de la realidad y en el "cálculo computacional" de los modelos matemáticos. Los algoritmos de IA, en tanto que se hallan exentos de sesgos, alimentan su "razonamiento" exclusivamente de datos, de los cuales extraen patrones comportamentales referentes y toman decisiones basados en estos. No obstante, la *razón algorítmica* se basa en una idea errónea y falaz. En primer lugar, tanto los datos de la realidad como el propio código fuente de los algoritmos están llenos de sesgos. Y en segundo lugar, el cálculo computacional ni puede interpretar y criticar lo dado, ni tampoco diferenciar entre lo vigente y lo válido. Por tanto, su "razonamiento" basado en patrones y no en procesos discursivos puede generar decisiones muy inmorales.

[5] Por *hiperética artificial* se entiende aquel enfoque de saber moral cuyos contenidos son generados por algoritmos de IA a partir del procesamiento computacional de los datos y metadatos en línea que genera la sociedad digitalmente hiperconectada.

[6] Por *dato* se entiende la unidad mínima de información (Hidalgo, 2015). Es decir, tiene que ver con la información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho" (DRAE, 2018). Por metadato se entiende la unidad mínima de información sobre un dato de algo.

Entre las distintas tareas de la ética, por consiguiente, destacan tres de ellas por encima de las demás. En primer lugar, la *clarificación conceptual* de todo aquello vinculado con el saber moral, como, por ejemplo, qué significan y qué relación guardan entre sí conceptos como ética, moral, libertad, dignidad, participación, agencia moral, bien común, virtud cívica, excelencia, principios, valores y normas morales, conciencia moral, etc. En segundo lugar, la *fundamentación de lo moral*; es decir, el proceso que permite dar razón de los principios, valores, normas, deberes, virtudes y móviles que conforman el marco prescriptivo de las relaciones humanas en cualquier momento y lugar. Y, en tercer lugar, la aplicación de todo el bagaje alcanzado en las primeras dos tareas descritas para que lo justo y lo *felicitante* acontezcan en las diferentes esferas funcionales de la sociedad; es decir, su introducción en los diferentes espacios de actividad para orientar la apropiación, aplicación y uso práctico que los seres humanos hacen desde la libertad de los principios, valores, normas, deberes y virtudes para proyectarse una vida buena en relación con los demás.

Siguiendo esta definición y estas tareas de la ética, actualmente coexisten diferentes corrientes filosóficas, como el aristotelismo, el utilitarismo y el deontologismo. Estas contribuyen independiente o conjuntamente a que lo justo y *felicitante* acontezca en la sociedad y sus distintas esferas funcionales y relacionales. No obstante, un enfoque deontológico de ética discursiva como el propuesto y desarrollado por Karl Otto Apel y Jürguen Habermas en la década de 1980 y 1990 (Apel, 1985; Habermas, 1984a, 1987; 1996) y madurado y ampliado por Adela Cortina, Jesús Conill y Domingo García-Marzá desde entonces (Cortina, García-Marzá y Conill, 2003; Cortina, 1986, 2007, 2010, 2017; 2021; Conill, 2006, 2019; García-Marzá, 1992, 2019), puede resultar sumamente fructífero para afrontar los desafíos actuales de las sociedades maduras –aquellas con un nivel de desarrollo moral posconvencional– (Conill, 2023; García-Marzá y Calvo, 2022). Por ejemplo, los problemas que subyacen a la irrupción y expansión de modelos de *sociedad digitalmente hiperconectada* cuyas esferas funcionales y relacionales están cada vez más colonizadas por algoritmos dotados de inteligencia artificial[7]. Especialmente, porque a través de una ética discursiva desarrollada –dialógica y cordial– es posible criticar y orientar el uso de la libertad de los implicados, así como abordar la reconstrucción y promoción tanto de una cultura o carácter a la altura de lo exigible y deseable por una *sociedad hiperdigital* moralmente plural como unas virtudes y unos afectos que lo canalicen, posibiliten y promuevan en la práctica.

Por un lado, este enfoque de ética deontológica de carácter discursivo ofrece un criterio de racionalidad desde el cual criticar y dar sentido tanto al conocimiento como al comportamiento: la aceptación por parte de todos los afectados de las consecuencias derivadas de una norma, acción o decisión aplicada o aplicable. Para ello, la propuesta se sustenta sobre procesos de diálogo y deliberación entre afectados que, tendentes al entendimiento sobre lo justo, se sustentan sobre dos principios morales básicos: uno *procedimental* (D) –sólo pueden pretender ser válidas aquellas normas, acciones o decisiones que, dentro de un discurso práctico

---

[7] Los algoritmos modernos están vinculados con la campo de la ingeniería computacional. Se pueden describir como "(...) a finite set of rules that gives a sequence of operations for solving a specific type of problema" (Angius, Primiero & Turner, 2021: 4) y tienen, al menos, cinco características básicas: finitud, precisión, entrada, salida y efectividad (Angius, Primiero & Turner, 2021; Rapaport, 2016; 2005).

con ciertas reglas argumentales y principios morales, puedan suscitar la aprobación de todos los afectados– y uno *universal* (U) –toda norma, acción o decisión que aspire a ser catalogada como moralmente válida debe satisfacer la condición de que las consecuencias previsiblemente derivadas de su aplicación actual o virtual puedan ser aceptadas por todos los afectados presentes o futuros. Además, estos dos principios demandan que el proceso de diálogo y deliberación tendente al entendimiento sobre lo justo se apoye en tres valores morales (García-Marzá, 1992): inclusión –todos los afectados presentes y futuros deben poder participar en aquellos procesos donde se delibera sobre las normas, acciones o decisiones aplicadas o aplicables que le afectan o le pueden afectar en el futuro–, igualdad –todos los afectados que participan en los procesos deben poder argumentar con las mismas condiciones de acceso a la información, de ausencia de presiones, de tiempo de palabra, etc.– y reciprocidad –todos los intereses de los participantes deben ser incluidos en el proceso y estar disponibles para su revisión crítica por parte del resto de interlocutores válidos–

Por otro lado, este enfoque deontológico de carácter discursivo se ha ido madurando y ampliando a través de la introducción de una dimisión cordial, experiencial, hermenéutica, axiológica, y prudencial que permite reducir el exceso de procedimentalismo de la propuesta original de Apel y Habermas y afrontar su aplicación sobre las diferentes esferas funcionales y relacionales (Cortina, 1986; Conill, 2006; García-Marzá, 1992). Al respecto, además de preocuparse por reconocer la validez moral de las normas universalmente vinculantes que regulan los espacios funcionales y relacionales ocupados por seres libres y comunicativamente vinculados, el enfoque también se abre a la toma en consideración de las preocupaciones por los principios, los valores, las virtudes, los fines, los móviles de la acción y los afectos que orientan, cohesionan y motivan los comportamientos de los agentes comunicativa y afectivamente vinculados.

Desde este enfoque ético de carácter deontologista, universalista, mínimo, procedimentalista, crítico, hermenéutico, cívico y cordial, las sociedades maduras pueden desentrañar los *presupuestos normativos* que subyacen a sus diferentes esferas funcionales y relacionales –el marco de referencia que permite justificarlos– para poder criticar, orientar y motivar racionalmente tanto los objetivos como los comportamientos implicados. Asimismo, en el diseño y puesta a punto de métodos de resolución de problemas y conflictos morales y en la concreción de aquellas virtudes cívicas que permiten promover un carácter excelente en cada ámbito de acción y relación. Y, finalmente, en el discernimiento de las herramientas, los mecanismos y las pautas que permiten su aplicación práctica y recreación fáctica.

Hoy, empero, esta transformación dialógica del saber moral se está viendo alterada por la constante intromisión de los modelos matemáticos en los procesos de racionalización con arreglo a sentido que las sociedades maduras hacen servir para llegar a acuerdos intersubjetivos sobre lo justo o correcto y, con ello, atenuar la vulnerabilidad humana frente a lo dado y el constante aumento de la complejidad de los contextos en los que se relaciona (Conill, 2023; García-Marzá y Calvo, 2022; Pérez-Zafrilla, 2021). Los resultados alcanzados por la inteligencia artificial en diferentes ámbitos de actividad han generado la falsa creencia de que es posible y deseable un enfoque hiper-artificial –*datificado*, hiperconectado y algoritmizado– de clarificación, fundamentación y aplicación de lo moral cuyos impactos ya se están dejando notar sobre los grupos más vulnerables de la sociedad. Es decir, se está promoviendo el paso de la transformación dialógica a la transformación digital

del ámbito moral para responder a las tres preguntas que, según Kant, se formula la razón: qué se puede conocer, qué se debe hacer y qué me cabe esperar (1968a, 1968b, A805, B833).

El interés que subyace a la transformación digital de lo moral no es nuevo. Responde, siguiendo a Habermas (1984b, 1987), al cientificismo e interés técnico de antaño por el control y dominio del mundo natural, pero también social. Un interés que hoy resurge con mayor fuerza y sutileza a través de la IA gracias a la supuesta objetividad y neutralidad que proporcionan sus algoritmos en cualquier esfera funcional y relacional de la sociedad, así como su capacidad de distorsión sobre el *mundo de la vida* (García-Marzá y Calvo, 2024). Mientras que las personas –parecen sugerir sus defensores (Matsumoto, 2020)– son portadoras de afectos, intereses, valores y principios subjetivos que introducen sesgos en los procesos de generación de conocimiento, las máquinas son portadores de modelos (matemáticos), datos y metadatos objetivos que ofrecen neutralidad en los procesos de construcción de saber sobre cualquier cosa de este mundo, como la ética. De este modo, el desarrollo social y humano queda reducido a un mero proceso de desarrollo científico-técnico. Se trata, tal y como advirtió Habermas, de (…) un progreso cuasi–autónomo de la ciencia y de la técnica, del que de hecho depende la otra variable más importante del sistema, es decir, el progreso económico" (Habermas, 1984b: 88), que produce fuertes anomalías en el mundo de la vida –como pérdida de sentido, anomia y patologías psicosociales– y pone en peligro la necesaria autonomía de sus elementos estructurales –cultura, sociedad y personalidad– y el desarrollo de sus procesos internos de racionalización[8] (Habermas, 1987: 279-280). En suma,

> (…) una perspectiva en la que la evolución del sistema social parece estar determinada por la lógica del progreso científico y técnico. La legalidad inmanente de este progreso es la que parece producir las coacciones materiales concretas a las que ha de ajustarse una política orientada a satisfacer necesidades funcionales. Y cuando esta apariencia se ha impuesto con eficacia, entonces el recurso propagandístico al papel de la ciencia y de la técnica puede explicar y legitimar por qué en las sociedades modernas ha perdido sus funciones una formación democrática de la voluntad política en relación con las cuestiones prácticas y puede ser sustituida por decisiones plebiscitarias relativas a los equipos alternativos de administradores (Habermas, 1984b, 88).

En este sentido, la *datificación* –la capacidad de la IA de reducir toda la realidad social y humana a datos y metadatos objetivos primero y de transformarlos en información relevante y conocimiento aplicable después–, se ha convertido en uno de los fenómenos principales del cientificismo e interés técnico por el control y dominio de la vida social y humana.

> Mediante los sistemas de inteligencia artificial los datos que se extraen de los usuarios de los medios digitalizados son aprovechados para lograr un conocimiento predictor y conformador de la conducta humana. El nuevo conocimiento de los datos permite lograr poder hegemónico sea a través del mercado o del Estado (...). Sea por la búsqueda del beneficio económico o del poder político y el control social, la inteligencia artificial se está utilizando para la apertura crítica de la razón pública. El nuevo sistema de acción

[8] Para un estudio pormenorizado de la teoría habermasiana, ver García-Marzá (1992).

> social produce una expropiación de los datos y una creciente dominación de las personas. Corren peligro la libertad personal, la privacidad y la intimidad (Conill, 2021).

Resulta especialmente relevante en este sentido la *etificación,* un subconjunto de la *datificación* que se preocupa por construir saber moral sintético de forma artificial (*sintetificación*) a través de la conversión de la realidad comportamental de los agentes morales en paquetes de datos y metadatos computables y su posterior almacenamiento, procesamiento, conversión y explotación mediante modelos matemáticos artificialmente inteligentes.

3. De la *datificación* a la *etificiación*: la transformación algorítmica de lo moral

El advenimiento de la *datificación*, promovida por algunas grandes empresas y agencias gubernamentales por el enorme potencial estratégico y predictivo que supuestamente atesora (Mayer-Schöenberger & Cukier, 2013; Saura, 2022), constituye uno de los efectos más visibles de la aplicación y convergencia sinérgica del Internet de las Cosas (IoT), el Big Data y la inteligencia artificial sobre la sociedad digitalmente hiperconectada.

La *datificación* es un proceso que se vale de los *ecosistemas ciberfisicos* y la ingente cantidad de datos y metadatos que estos generan para, a través de modelos matemáticos artificialmente inteligentes, rastrear, segmentar y/o encontrar patrones comportamentales en un ámbito, colectivo o sociedad concreta o general con la finalidad de satisfacer un objetivo estratégico, como puede ser la seguridad y la educación de la ciudadanía por parte de un gobierno o la innovación y la personalización de nuevos productos y servicios por parte de una organización económica.

El éxito de la *datificación* y su rápida asimilación en toda esfera de actividad humana, ha tenido mucho que ver con la visión inocente, neutra y edulcorada que han proyectado sobre la sociedad distintas organizaciones gubernamentales y económicas. Al respecto, los discursos suelen destacar los beneficios que ello puede producir para la sociedad, como una mayor seguridad ante ataques terroristas, un mayor control de las pandemias, una mejora sustancial en el diagnóstico y cura de enfermedades, una gestión más eficiente de los procesos productivos, un aumento considerable de la eficacia de los medicamentos, un uso más sostenible de los recursos escasos e incluso una toma de decisiones más objetiva, inclusiva y justa. De ahí que la *datificación* sea vista por algunos sectores de la sociedad como "(…) un medio legítimo para acceder, comprender y monitorizar el comportamiento de las personas" (van Dijck, 2014), puesto que los posibles usos instrumentales de su aplicación sobre la sociedad quedan como un mal menor frente a los grandes beneficios que estos generan para la sociedad en general. Es, podríamos decir, la versión más moderna, compleja y a la vez sutil de la metáfora de la *mano invisible* elucubrada por Bernard Mandeville y Adam Smith en el siglo XVIII; una *hiperestésica mano digital* que mueve implacablemente los hilos de los nuevos enfoques hipermasivos de economía, pedagogía y democracia, entre otros. Como argumenta Conill,

> La creciente extensión de las tecnologías de la información y de la digitalización en todos los ámbitos de la vida humana, extrayendo una enorme cantidad de datos y ampliando

> el conocimiento del comportamiento de la gente, genera un poder que es empleado para proyectos comerciales y/o políticos cada vez más voraces. No sirve para fomentar la razón pública, sino para predecir y conformar la conducta. En vez de empoderar a los ciudadanos, genera dependencia y adicción, activando el interés de dominio y ampliando la capacidad de ejercer la dominación de modo efectivo en la vida social. Se pone en peligro el ejercicio de la razón y de la libertad, atentando contra las bases de la concepción ilustrada del hombre y de la sociedad, así como contra la democracia constitucional (Conill, 2021: 44).

Entre los casos más relevantes al respecto, destaca el programa de vigilancia masiva puesto en marcha por las autoridades chinas para prevenir, controlar y/o mejorar el comportamiento de la ciudadanía en diferentes espacios de actividad[9]. Por un lado, destaca el programa *crédito social*, donde millones de cámaras dotadas de tecnología de reconocimiento facial colocadas en espacios públicos y privados, como calles, metros y casas de alquiler, son utilizadas por un algoritmo llamado *skynet* para recabar información comportamental, medir el grado de civismo de la ciudadanía digitalmente *hiperconectada*, y elaborar una tarjeta personalizada de buen ciudadano que da o impide el acceso a las plazas, contratos o ayudas estatales (Laniuk, 2021). Por otro lado, también destaca el Sistema Inteligente de Gestión del Comportamiento en las Aulas, donde cámaras dotadas de tecnología de reconocimiento facial y colocadas sobre las pizarras, envían información a un algoritmo para que evalúe el grado de atención del alumnado, clasifique su estado emocional en cada momento (neutro, feliz, triste, decepcionado, molesto, asustado o sorprendido) y avise al profesorado cuando identifique anomalías como la falta de atención o el acontecer de emociones negativas (Brown, Statman & Sui, 2021).

También es significativo el uso de modelos matemáticos artificialmente inteligentes por parte de organismos judiciales y policiales en todo el mundo para ejecutar sentencias, aplicar penas y predecir delitos, delincuentes y víctimas (D'amato, Silver, Newsome & Latessa, 2020). Al respecto, destaca el uso cada vez mayor de programas informáticos como *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS), cuyo algoritmo realiza cálculos estadísticos basados en los metadatos en línea y el historial comportamental de los delincuentes para aconsejar a jueces/as sobre la culpabilidad de un acusado/a, el número de años y el tipo de pena que se debe imponer a un/a delincuente o el grado de reincidencia de un/a preso/a con posibilidad de optar a la libertad condicional. Por otro lado, también destaca el uso de programas predictivos como *National Data Analytics Solution* (NDAS), cuyo algoritmo efectúa cálculos estadísticos basados en los metadatos en línea y el historial comportamental de los/as habitantes de un barrio o una ciudad concreta para ofrecer a la policía predicciones sobre futuros delitos, malhechores y víctimas.

Finalmente, cabe destacar el uso de algoritmos dotados de inteligencia artificial para la creación, desarrollo, venta y distribución de productos y servicios. Amazon, por ejemplo, ha desarrollado un algoritmo de IA que explota los datos y metadatos en

---

[9] La vigilancia masiva está muy extendida por todo el mundo. La gran mayoría de países, también España, ha comprado tecnología china de reconocimiento facial para el control comportamental de la ciudadanía. Al respecto, el estudio AI Global Surveillance Index (Feldstein, 2019) ofrece datos sobre 176 países que disponen de esta tecnología.

línea y el historial de búsqueda y compra de un cliente concreto para predecir aquello que le pedirá en el futuro y enviárselo antes de que sepa que lo quiere comprar. Otro ejemplo destacable son las empresas que, como Lantia Publishing, ofrecen modelos matemáticos artificialmente inteligentes capaces de monitorizar y procesar datos comportamentales de los usuarios del sector editorial y cinematográfico para detectar las tendencias del mercado y captar y desarrollar productos de éxito.

Los supuestos logros alcanzados por la implementación de procesos de *datificación* de la realidad social en distintas esferas de actividad humana –como la política, la economía, la pedagogía o la salud– han despertado el interés de académicos, profesionales y tecnólogos por la transformación digitalmente del ámbito moral. Esta preferencia cada vez más extendida se conoce con el nombre de *etificación* (Calvo, 2019, 2021; García-Marzá y Calvo, 2024; Conill, 2021)–. Se trata de un subconjunto del proceso de *datificación* que consiste en el intento de convertir los hechos, las opiniones, las preferencias, los hábitos y las conductas morales de la ciudadanía digitalmente *hiperconectada* en datos y metadatos en línea para, desde un enfoque descriptivo –lo que es– y no normativo –lo que debería ser– y basado en un criterio de mayorías[10] –el mayor bien para la mayor cantidad de gente posible–, "(…) establecer mediante modelos matemáticos artificialmente inteligentes qué es moralmente deseable –*felicitante*– y válido –justo– para la sociedad" (Calvo, 2019).

Entre las principales manifestaciones del interés por la *etificación* y su aplicación práctica, destacan los diferentes intentos por dotar los modelos matemáticos artificialmente inteligentes de competencias y capacidades morales a través de la recopilación de datos masivos sobre el comportamiento moral de la ciudadanía. Un claro ejemplo de ello son los experimentos *The Moral Machine*, una plataforma desarrollada por el MediaLab del MIT para recabar datos sobre situaciones moralmente dilemáticas cuando se conduce un vehículo, y *Moral Choice Machine*, una investigación para intentar dotar a las máquinas de las competencias necesarias para "(…) aprender a tomar elecciones éticas o incluso morales" mediante la extracción, cuantificación y comparación de fuentes culturales sobre elecciones morales (Jentzsch et al, 2019). Tras estos y otros proyectos y estudios similares subyace la falsa creencia de que es posible para una máquina dotada de IA establecer qué es justo y deseable en un contexto concreto mediante la cuantificación de opiniones, comportamientos y magnitudes morales dilemáticas.

Otra clara manifestación del actual interés por la *etificiación* y su aplicación práctica, son los intentos por desarrollar algoritmos de IA que puedan tomar decisiones justas basadas en la objetividad de los datos y los metadatos en línea (Calvo, 2019, Conill, 2023). Destaca al respecto la emergencia de diferentes casos de políticos algorítmicos alrededor del mundo. Por un lado Alisa, la inteligencia artificial que la tecnológica Yandex intentó presentar a las elecciones rusas de 2018 alegando que esta era más justa que los políticos de carne y hueso, puesto que no se deja llevar por las emociones, no busca ventajas personales y no emite juicios sobre los intereses y las preferencias de la ciudadanía. Por otro Por otro, IA Mayor, la inteligencia artificial que Michihito Matsuda –fundador de la empresa extinta File Rogue– presentó en 2018 a la alcaldía de Tama New Town (Japón) para acabar con la corrupción política, dialogar y llegar a entenderse con el resto de fuerzas

[10] Recolección y suma de hechos, experiencias, conductas y respuestas dadas ante situaciones diversas, etc. que se convierten en datos que se utilizan para tomar decisiones.

políticas por el bien del distrito y proporcionar "(…) fair and balanced opportunities for everyone" (Johnston, 2018). También SAM, la inteligencia artificial que el grupo tecnológico compuesto por Walter Langelaar, Nick Gerritsen y Walter Langelaar pretendió presentar a las elecciones presidenciales neozelandesa con el pretexto de que esta, a diferencia de los políticos de carne y hueso, será capaz de tomar decisiones justas por el bien común de la ciudadanía. Asimismo, el *Partido sintético* danés (*Det Syntetiske Parti*) que, lanzado en agosto de 2022 por el colectivo de artistas Computer Lars y compuesto únicamente por algoritmos artificialmente inteligentes, se presentó a las elecciones generales de 2023 para elevar "las visiones políticas de la persona común" y, de ese modo, mejorar la democracia (García-Marzá y Calvo, 2022). Y, finalmente, ION, la primera inteligencia artificial nombrada oficialmente como consejera gubernamental de un país democrático (Rumanía) para, entre otras cosas, asesorar al primer ministro sobre acciones y decisiones políticas (García-Marzá y Calvo, 2024). En suma, tal y como argumenta Matsumoto en *The Day AI Becomes God: The Singularity Will Save Humanity*,

> Lo siguiente quizás podría describir una implementación ideal de democracia: haga un uso completo de la inteligencia artificial para determinar primero el sentimiento público (insatisfacción con el *statu quo* y demás), identifique las ventajas y desventajas a largo plazo de varias opciones de políticas disponibles, eduque a la población explicándoles de una manera fácil de entender, una vez más medir el sentimiento público y luego decidir e implementar políticas de acuerdo con los hallazgos (Matsumoto, 2018, 161).

Finalmente, también cabe destacar como manifestación del interés por la *etificiación* y su aplicación práctica, los intentos por incluir algoritmos artificialmente inteligentes en los procesos de deliberación moral en el ámbito asistencial. Es significativo el caso de ORACLE (*Organizer of the Rational Approach in Computational Learning Ethics*), un modelo matemático artificialmente inteligente diseñado para ayudar a los comités de bioética en la resolución de conflictos y la toma de decisiones de carácter ético en la práctica clínica (Motta et al., 2016). Otro ejemplo de este fenómeno incipiente es el trabajo "Investigation of the visual attention role in clinical bioethics decision-making using machine learning algorithms" (Fernandes et al. 2017), donde se propone el uso de un algoritmo de ML y dos técnicas de *data mining* para, utilizando datos de seguimiento ocular, elaborar un modelo predictivo que sirva de ayuda en los procesos de toma de decisiones en bioética clínica, sobre todo en aquellos vinculados con la eutanasia.

No obstante, tras este interés por la *datificación* de la realidad social y moral subyace una ideología científico-técnica y una imposición sistémica –económico-estatal– que produce trastornos y anomalías importantes en el correcto desarrollo de la sociedad digitalmente *hiperconectada* del siglo XXI. La intromisión de los algoritmos de IA en aquellos procesos de racionalización que pretenden dotar de sentido diferentes cosas del mundo objetivo, social y subjetivo, como la generación de conocimiento, el accionar humano, la construcción de la personalidad o la proyección de una vida buena, está produciendo una sutil pero implacable colonización del mundo de la vida que, a través de imperativos sistémicos que está alterando de forma radical las estructuras y el correcto funcionamiento y desarrollo de las sociedades modernas en todas sus dimensiones.

> (...) mediante la datificación se impone una concepción utilitarista de la vida y de la sociedad, reforzada por los nuevos instrumentos tecnológicos y el tipo de conocimiento que ofrecen, orientado primordialmente por el cálculo. De hecho, los procesos de "datatificación" o "datificación" van invadiendo incluso el ámbito moral produciendo una "etificación". Se relega o anula el diálogo y la reflexión crítica regida por las pretensiones de validez, en favor de la agregación de las opiniones, preferencias y hábitos. Por tanto, se desconsidera o elimina la razón comunicativa en favor del cálculo y la matematización de los datos, que sirve para predecir el comportamiento y alimentar la razón instrumental y estratégica. Se elimina así la crítica recíproca entre los que participan en la esfera pública, porque en el fondo se presupone una concepción de la sociedad orientada por un individualismo cuantitativista (Conill, en prensa: 11).

El impacto más preocupante de todo ello, es la falsa creencia de que es posible y deseable un modelo ético de clarificación, fundamentación y aplicación basado en la *datificación*, la *algoritmización* y la *hiperconectividad digital.* Entre otras cosas, porque el enfoque hiper-artificial resultante puede servir para legitimar las injusticias sociales, para tergiversar las normas, valores, sentimientos, virtudes y pautas moralmente válidas, para derivar la responsabilidad de las acciones y decisiones en *agentes amorales*, y para desplazar el posible consenso entre afectados por un análisis computacional de la información subjetiva disponible.

4. Hiperética artificial: la colonización algorítmica del mundo de la vida

Desde una perspectiva crítica, Habermas (1984a, 1987) propone que la sociedad moderna se halla estructurada sobre dos niveles de aplicación autónomos, pero complementarios y necesarios: el funcional del mundo sistémico y comunicativo del subsistema del mundo de la vida. Por un lado, el *nivel sistémico* comprende la esfera del estado y del mercado, que dan cuenta del mundo político administrativo y mercantil y cuya proyección y desarrollo depende de la puesta en marcha de acciones estratégicas, es decir, de procesos de racionalización con arreglo a fines. Por otro lado, el *nivel subsistémico* del mundo de la vida está conformado por tres esferas –la cultura, la sociedad y la personalidad– que dan cuenta del mundo objetivo –en tanto que totalidad de las entidades sobre las que son posibles enunciados verdaderos–, mundo social –en tanto que totalidad de las relaciones interpersonales legítimamente reguladas– y mundo subjetivo –en tanto que totalidad de las propias experiencias vividas por los sujetos de la sociedad–, cuya proyección y desarrollo depende de la implementación de acciones comunicativas, es decir, de procesos internos de racionalización mediados por el entendimiento intersubjetivo.

El mundo de la vida, que Habermas define como "(…) un acervo de patrones de interpretación transmitidos culturalmente y organizados lingüísticamente" (1987), representa el conjunto de convicciones y/o autoevidencias incuestionadas de los cuales hacen uso los sujetos dotados de habla y acción para introducirse en aquellos *procesos cooperativos de interpretación* cuyo principal objetivo es entenderse sobre diferentes cosas de este mundo. Es decir, desde, y a partir de, el mundo de la vida que le es común al hablante y al oyente, ambos buscan entenderse sobre algo en el mundo objetivo –verdad–, en el mundo social –justicia– y en el mundo subjetivo –veracidad– (1987).

En tanto que al mundo de la vida le corresponde el entendimiento como tal, y, por tanto, la definición del sistema social en su conjunto (Habermas, 1987), los distintos mecanismos de coordinación del sistema –mercado y estado– se ven abocados a anclar sus raíces en el mundo de la vida para dotarse del sentido y la legitimidad que justifique su existencia y permite su perdurabilidad (Habermas, 1987). Un gobierno o un mercado que dé la espalda al mundo de la vida corre el peligro de quedar desdibujado tras déficits de sentido que ponen en jaque su supervivencia y desarrollo. No obstante, el aumento constante de la complejidad del mundo de la vida durante la modernidad y la contemporaneidad, fruto de la emergencia de nuevas y mayores exigencias de sentido, dieron pie a un proceso de colonización de sus diferentes esferas –cultura, sociedad y personalidad– mediante imperativos sistémicos que, como el poder y el dinero, producen anomalías – anomia, alienación y psicopatologías– que ponen en peligro su autonomía y, por ende, su desarrollo y subsistencia.

> Cuando se introduce la violencia (Gewalt) como alternativa al mecanismo de coordinación de la acción que representa el entendimiento y el poder (Macht) como producto de la acción orientada al entendimiento, se obtiene, además, la ventaja de no perder de vista las formas de ejercicio indirecto de la violencia que hoy predominan. Me refiero a esa violencia patógena que inadvertidamente penetra en los poros de la práctica comunicativa cotidiana y puede desplegar en ella su latente eficacia en la medida en que el mundo de la vida queda entregado a los imperativos de subsistemas funcionales autonomizados y cosificados por las sendas de una racionalización unilateral (Habermas, 1987).

En la actualidad, este hecho se ha visto exponencialmente agravado por los procesos de transformación digital. El nuevo mundo *hiperdataficado* que le subyace, caracterizado, por su alta dependencia de las Tecnologías de la Información y la Comunicación (TIC) (Floridi, 2012), la desfiguración y deshumanización de las relaciones sociales (Donati, 2019), la conversión de realidad social en datos y metadatos computables (Mayer-Schöenberger y Cukier, 2013), la reducción del ser humano a un mero terminal de flujo de datos y metadatos en línea (Zuboff, 2019), la normalización de niveles insostenibles de entropía, desincronización y aceleración de la vida social (López-González, 2022), y la supremacía del valor de lo virtual frente a lo real (Calvo, 2023), ha aumentado la complejidad del mundo de la vida por la emergencia de nuevas y masivas necesidades de sentido. Este hecho ha generado que la colonización sistémica del mundo de la vida venga determinada por una cada vez mayor tecnificación e instrumentalización de sus componentes estructurales a través de procesos de *datafícación* de la producción simbólica en todas sus dimensiones. El resultado de este mundo *hiperdataficado* es la imposición de un punto de vista sesgado donde el desarrollo del sistema social parece estar determinado por la lógica del progreso científico-técnico (Habermas, 1984b) y donde los procesos de toma de decisiones vinculados se reducen al mero cálculo computacional del conjunto de datos y metadatos disponibles.

Siguiendo a Habermas, el problema principal de este mundo cada vez más gobernado por algoritmos de inteligencia artificial y, por tanto, cada vez más dependiente de los procesos de *datafícción* e *hipeconectividad digital*, es que el cálculo llevado al extremo –principal característica de los procesos de toma de decisiones de este tipo de perspectivas científico-técnicas– "(…) deja en estado de

pureza lo que son decisiones, es decir, las limpia de todos aquellos elementos que aún podían considerarse accesibles a algún análisis de tipo vinculante" (Habermas, 1984b). De este modo, la *datificación* forja un proceso acrítico y anómicamente deficiente de evolución del sistema social que produce graves interferencias en la reproducción simbólica, en la generación de sentido y en la coordinación de la acción.

Un indicador claro de todo ello, es el incipiente interés de un número cada vez más grande e influyente de académicos, tecnólogos y simpatizantes por concretar y aplicar un modelo *algoritmizado* de clarificación, fundamentación y/o aplicación de lo moral: la *hiperética artificial (artificial hyperethics* en inglés), la cual produce graves anomalías en los procesos de racionalización con respecto a sentido en el mundo objetivo, social y personal. Esta, entendida como una forma *digitalizada* de captar y generar saber moral de forma *amplia* y *objetiva* mediante el uso de las TIC y la puesta en marcha de procesos de *datificación* del ámbito moral –*etificación*–, pretende identificar en tiempo real qué acciones, normas y decisiones son moralmente justas y *felicitantes* para la sociedad, qué virtudes están detrás de un carácter excelente de la ciudadanía, qué sentimientos e intereses constituyen motivaciones para la acción moral de las sociedades y qué competencias y capacidades permite su recreación fáctica, entre otras cosas.

Así, por un lado, el prefijo *hiper* sirve para identificar este enfoque artificial de ética en un cuádruple sentido. En primer lugar, como *hiperbólico*, en tanto que expectativas en grado superior a las humanas sobre las capacidades morales y los resultados que generan o pueden llegar a generar los modelos artificialmente inteligentes en el ámbito moral. En segundo lugar, como *hipertextualidad* e *hiperconectividad*, en tanto que conjunto estructurado de *cosas* –también personas– unidas entre sí por enlaces, conexiones y datos. En tercer lugar, como *hiperinformación* e *hiperhistoria,* en tanto que hechos y procesos sociales altamente dependientes de las TIC y sus resultados. Finalmente, como *hiperespacio,* en tanto que representación de una *esfera pública virtualizada* que, hipotéticamente, va más allá de las tres dimensiones espaciales conocidas. Mientras que, por otro lado, el adjetivo *artificial* sirve para identificar este enfoque de ética con un constructo computacional; es decir, como el corolario resultante de la aplicación de diversas técnicas y tecnologías informáticas de emulación del razonamiento, el aprendizaje y el juicio moral de los seres humanos.

Al respecto, entre la literatura especializada existen numerosos estudios y experimentos (Allen, Wallach & Franklin, 2011, 2009; Awad et al., 2018; Bostrom, 2014; Floridi & Sanders, 2004; Hooker, 2018; Jentzsch et al, 2019; Matsumoto, 2018; Motta et al., 2016; Siqueira-Batista, 2014; Tigard, 2021[11]) que contribuyen consciente o inconscientemente a la promoción y desarrollo de la *hiperética artificial* y de los cuales subyacen ciertos conceptos y características básicas. A saber:

1. Sentido moral artificial. En primer lugar, la *hiperética artificial* desplaza aquellos procesos deliberativos donde los afectados por una norma, acción o decisión aplicada o aplicable intentan entenderse sobre la validez moral de algo en el mundo a través del diálogo y el posible acuerdo intersubjetivo, por procesos de análisis computacional de grandes bases de datos y metadatos

[11] Sobre este tema, existen muchos más ejemplos en la literatura especializada. Aquí sólo se ofrece una pequeña y selectiva muestra de todos ellos. Una revisión crítica y exhaustiva de los diferentes estudios que alimentan la *hiperética artificial* requiere de un trabajo individualizado.

mediante modelos matemáticos artificialmente inteligentes. La inteligencia artificial capta el *sentido moral* de las cosas del mundo a través del cómputo de los datos y metadatos disponibles. A través de ello, la inteligencia artificial extrae conjuntos de ítems, atributos o magnitudes morales recurrentes; es decir, patrones morales que luego utiliza para orientar a la sociedad sobre lo justo y *felicitante* o para utilizar ella misma como criterio de racionalidad en los procesos de toma de decisiones donde participa. El sentido moral, por tanto, queda reducido a una descripción y agrupación de la realidad comportamental de la sociedad; es decir, a un agregado de voluntades, comportamientos y opiniones en línea.

2. Agencia moral artificial. En segundo lugar, la *hiperética artificial* extiende los límites de la agencia moral para poder incluir ciertas entidades amorales, como los modelos matemáticos dotados de inteligencia artificial. En el estudio "On the Morality of Artificial Agents" (2004), por ejemplo, Luciano Floridi y J.W. Sanders defienden la moralidad de aquellos algoritmos dotados de inteligencia artificial que toman decisiones. Para ello, aplican un concepto de agencia moral que elude cuestiones como la libertad, los estados mentales y la responsabilidad del agente y un discurso que subyuga la ética al derecho y reduce la libertad a un mero proceso de libre albedrío, entre otras muchas cosas[12].
3. Condicionalidad moral artificial. En tercer lugar, la *hiperética artificial* mantiene una total dependencia de las TIC y del flujo de datos y metadatos masivos. La desconexión o mal funcionamiento de las TIC significa la parálisis del enfoque por la falta o escasez de datos procesables y convertibles en información relevante y conocimiento aplicable o por la deformación y/o arbitrariedad de sus resultados. La *conexión* o *desconexión moral* parcial o general de este tipo de enfoque, por tanto, está condicionada y/o influida por las decisiones estratégicas de las grandes tecnológicas, puesto que son estas las que controlan el flujo de datos y metadatos en línea, custodian la mayor parte de las grandes bases de datos y conceden acceso a los distintos niveles de hiperconectividad que permiten influir en mayor o menor medida en los algoritmos de inteligencia artificial implicados.
4. Objetividad moral artificial. En cuarto lugar, la *hiperética artificial* formula una propuesta descriptiva de clarificación, fundamentación y aplicación de lo moral. Este enfoque *dataficado*, hiperconectado y algoritmizado de ética no pretende prescribir la realidad social para encauzarla y mejorarla, sino describir los hechos y las opiniones, las costumbres y los comportamientos de la *ciudadanía digitalmente hiperconectada* para obtener patrones estadísticos que permitan determinar qué decisiones, pautas y conductas son *objetivamente* justas y *felicitantes*.
5. Realidad moral artificial. Y, en quinto lugar, la *hiperética artificial* promueve una estructura binaria de la realidad para la resolución de la conflictividad, la coordinación de la acción y la toma de decisiones. Basada en dilemas morales y centrada en los resultados de las acciones y decisiones, la propuesta reduce toda la realidad a dos únicas posibilidades para salvar la alta complejidad de

[12] Este tipo de perspectivas han fomentado el uso actual de algoritmos de IA en los procesos de toma de decisiones en las diferentes esferas de actividad social para eludir la responsabilidad.

> la realidad. De esta forma, tal y como se puede comprobar en experimentos como *The Moral Machine* (Awad et al., 2018), ante un hecho concreto este enfoque ofrece dos posibilidades excluyentes –la elección de una de ellas exige el olvido de la otra– y una solución *racional* –basada en los datos y metadatos disponibles y el cálculo de las consecuencias previsibles–[13] que, en la mayoría de los casos, suele ser inmoral[14].

Como resultado de todo ello, la *hiperética artificial* convierte el saber moral –prescriptivo y emancipatorio– en científico-técnico –predictivo y determinista–. Este hecho genera interferencias en la concreción de aquellas normas, valores, sentimientos, virtudes y pautas moralmente válidas y *felicitantes*, produciendo déficits de legitimidad, escasez de normas vinculantes, inopia axiológica, podredumbre virtuosa y desafección participativa que pone en peligro la racionalidad, solidaridad y autonomía del mundo de la vida (Habermas, 1987). Todo ello hace de la *hiperética artificial* un mecanismo apto para la legitimación de las injusticias sociales y la pervivencia de pautas y comportamientos moralmente indeseables para las sociedades con un nivel posconvencional de desarrollo moral que limitan el progreso de sus diferentes esferas. Como advirtió Habermas, es la legitimidad inmanente de este tipo de procesos, "(…) la que parece producir las coacciones materiales concretas a las que ha de ajustarse una política orientada a satisfacer necesidades funcionales. Y cuando esta apariencia se ha impuesto con eficacia, entonces el recurso propagandístico al papel de la ciencia y de la técnica puede explicar y legitimar por qué en las sociedades modernas ha perdido sus funciones una formación democrática de la voluntad política en relación con las cuestiones prácticas y puede ser sustituida por decisiones plebiscitarias relativas a los equipos alternativos de administradores" (Habermas, 1984b).

5. Crítica al diseño y aplicación de un enfoque hiper-artificial de ética

Desde una perspectiva ética de carácter universalista, deontológica, procedimentalista, mínima, hermenéutica, dialógica, cordial y crítica como la aquí defendida, el saber moral no puede reducirse a la agregación y procesamiento computacional de datos y metadatos en línea sobre los intereses, opiniones y comportamientos de la sociedad digitalmente hiperconectada. Especialmente, porque de esta forma se desprecia su autonomía, la capacidad de la ciudadanía de criticar lo dado, de deliberar sobre los mejores cursos de acción para la resolución de conflictos, de llegar a acuerdos, coordinar de la acción, de convencer mediante buenas razones sobre la *ligatio* que *ob-liga* a todo el mundo a actuar de una determinada forma, y de cambiar para mejorar, revisar o erradicar diferentes cosas y comportamientos aun cuando ello vaya en contra del propio interés.

---

[13] Como contraargumentan Barry Dewitt, Baruch Fischhoff y Nils-Eric Sahlin, "The 'moral machine' experiment for autonomous vehicles devised by Edmond Awad and colleagues is not a sound starting place for incorporating public concerns into policymaking" (Dewitt, Fischhoff & Sahlin, 2019).

[14] En buena parte de las ocasiones, el experimento Moral Machine solicita al jugador que tome una decisión dilemática basada en la cantidad, entre matar a una y más personas, o en la discriminación, entre matar a una persona joven o anciana (edadísmo), una mujer o un hombre (misoginia).

Existe, por consiguiente, un vínculo estrecho e inalienable entre saber moral y participación ciudadana, y un enfoque *hiper-artificial* de clarificación, fundamentación y aplicación de lo moral que ataca los procesos de racionalización con arreglo a sentido, aquellos que permiten legitimar los comportamientos libres y los impactos que producen, representa un desafío de primer orden que exige ser abordado con seriedad y rigurosidad.

Al respecto, la aceptación y aplicación acrítica de un enfoque *datificado*, algoritmizado e *hiperconectado* de clarificación, fundamentación y aplicación de lo moral puede generar consecuencias inintencionadas y/o intencionadas sobre la sociedad y la ciudadanía, y, por ende, sobre las democracias maduras que basan su emergencia y desarrollo en la participación de los afectados por sus impactos. Desde mi punto de vista, destacan especialmente cuatro consecuencias: reificación, desigualdad, heteronomía y convencionalismo.

1. Reificación (*Verdinglichung*): la *hiperética artificial* es un enfoque que reduce al ser humano a un mero terminal de flujo de datos y metadatos masivos en línea. Para la *hiperética artificial,* el ser humano no parece ser más que una *cosa hiperconectada* dentro de un *ecosistema ciberfísico*, por lo que su valor ya no se ve como intrínseco, vinculado con su capacidad para autolegislarse, sino extrínseco, vinculado con su capacidad para estar hiperconectado y generar grandes volúmenes de datos y metadados en línea con los que alimentar los modelos matemáticos artificialmente inteligentes. La *hiperética artificial*, por tanto, se muestra como un enfoque falaz que promueve al ser humano como producto[15]–cuyo valor queda condicionado por el volumen de datos que aporte al sistema– y no como fin en sí mismo –cuya dignidad es incondicional e incondicionada–.
2. Desigualdad: la *hiperética artificial* es un enfoque que aumenta la complejidad social y, por consiguiente, la brecha de las desigualdades sociales en todas sus dimensiones. Su dependencia de las TIC y las grandes desigualdades existentes en el acceso y uso de tales tecnologías por parte de la ciudadanía, así como el sesgo misógino, aporófobo, homófobo y xenófobo que portan los datos y metadatos masivos en línea que utiliza para sus objetivos, hace de la *hiperética* un mecanismo maleable y propicio para generar descripciones sesgadas y/o manipuladas del saber moral.
3. Heteronomía: la *hiperética artificial* es un enfoque cuya apuesta por procesos artificiales de discernimiento y toma de decisiones sobre aspectos morales aboca al ser humano a una preocupante desafección participativa y un desinterés por las injusticias. Conforme va calando en la sociedad la falsa idea de que los algoritmos de IA están más capacitados que las personas para discernir aquello que es válido y deseable en una sociedad concreta o general, se va formando una peligrosa tendencia hacia la adopción de pautas paternalistas de comportamiento en las cuales la inteligencia artificial detenta todo el poder de decisión y los seres humanos acatan acríticamente sus dictámenes. Esta tendencia promueve un retroceso en el proyecto sociopolítico de la modernidad, donde la mayoría de edad intelectual del ser humano

[15] Los enfoques de ética aplicada como la ética tecnológica y ética digital no promueven este tipo de enfoque *hiperético* de clarificación, fundamentación y aplicación de lo moral.

constituye la piedra angular de su emergencia y desarrollo. En el nuevo proyecto sociopolítico que parece emanar de las *sociedades digitalmente hiperconectadas* parece entreverse un gobierno algorítmico de los procesos de racionalización con respecto a fines y, lo que es más preocupante, con respecto a sentido.

4. Convencionalismo: la *hiperética artificial* es un enfoque cuyo convencionalismo lastra las posibilidades de transformar la realidad social en un sentido justo y *felicitante*. Los modelos matemáticos dotados de inteligencia artificial carecen de las competencias necesarias para criticar lo dado y diferenciar entre la vigencia y la validez moral de una norma, acción o decisión. Por un lado, esta incompetencia de los algoritmos de inteligencia artificial impide el establecimiento y proyección de un mejor mundo posible para una sociedad concreta o general. Y, por otro lado, tal incompetencia algorítmica fomenta la continuidad de pautas y comportamientos anacrónicos e inaceptables para una sociedad madura con un nivel posconvencional de desarrollo moral.

6. Conclusiones

Estas y otras cuestiones son un desafío para la sociedad digitalmente hiperconectada y, por ende, para las sociedades maduras –plurales– con un nivel posconvencional de desarrollo moral. La transformación digital parece irreversible, y sus beneficios pueden ser altos para las sociedades. Sin embargo, su emergencia y desarrollo exige crítica constante y pautas adecuadas que eviten un incremento de la complejidad y, en consecuencia, de la vulnerabilidad de las personas, especialmente de aquellas pertenecientes a los colectivos más frágiles de la sociedad. Por ello, como argumento Conill (2022: 45).

> Hay que despertar del nuevo sueño dogmático con apariencia progresista y optar radicalmente por la libertad sin dejarse seducir ni someter como rebaños por las actuales coerciones del nuevo poder social de carácter científico-técnico. Hay que abandonar el tecno-determinismo optimista, porque las tecnologías digitales crearon expectativas que no sólo no se han cumplido, sino que han empeorado la situación para el desarrollo de la auténtica democracia.

Para ello, es necesario orientar la praxis digital en un sentido justo y responsable a través del diseño y aplicación de marcos de referencia adecuados –como directrices y códigos de ética, conducta, deontología y buenas prácticas–, establecer mecanismos para la deliberación y el diálogo de los afectados por sus consecuencias –como comités de ética–, implantar instrumentos que recaben información sobre el cumplimiento de los compromisos alcanzados –como las líneas éticas–, poner en marcha herramientas de rendición de cuentas –como los informes de explicabilidad– y promover una cultura adecuada basada en el seguimiento, aplicación e implementación de los principios éticos para una IA confiable (García-Marzá, 2017; García-Marzá y Calvo, 2024; High-level expert Group on Artificial Intelligence, 2019).

No obstante, cabe tener en cuenta las diferentes condiciones y perspectivas actuales y virtuales de los países que diseñan, aplican y utilizan IA. Como se ha

podido apreciar, en el trasfondo de este estudio subyace una perspectiva europea de IA confiable, basada en la orientación que proporcionan los 4 principios de las directrices éticas elaboradas por High-level expert Group on Artificial Intelligence (2019) –no-maleficencia, autonomía, justicia y explicabilidad– y sus 7 requisitos de aplicación para su recreación fáctica –Acción y supervisión humana, Robustez técnica y seguridad, Privacidad y gobernanza de datos, Transparencia, Diversidad, no discriminación y equidad, Bienestar social y ambiental y Responsabilidad–. No obstante, si bien países como Estados Unidos, China, Japón y Rusia no difieren en el trasfondo –la necesidad de orientar de manera justa y *felicitante* el diseño, aplicación y uso de la IA–, sí lo hacen en el resultado –los principios y valores[16]– y su aplicación –las políticas y estrategias–.

Por ello, para avanzar en la pretensión de universalidad y racionalidad que persigue una perspectiva de ética discursiva como la defendida en este estudio, cabe seguir realizando esfuerzos en la búsqueda de consensos intersubjetivos entre los y las afectados y afectadas sobre las consecuencias derivadas de la transformación digital de lo moral. Y para ello es necesario incrementar los esfuerzos en el diseño, aplicación e implementación de espacios de comunicación y *relacionalidad* que permitan captar en mayor medida la idiosincrasia, las perspectivas y los intereses de los distintos países implicados en el diseño, aplicación y uso de inteligencia artificial.

7. Referencias bibliográficas

Angius, Nicola, Primiero Giuseppe & Turner, Raymond (2021). "The Philosophy of Computer Science", The Stanford Encyclopedia of Philosophy. In: Edward N. Zalta (ed.), *Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University: Spring, pp. 1-21.

Aramayo, Roberto R. (1986). "Estudio preliminar", en Kant, Immanuel, *Teoría y Práctica*. Madrid: Tecnos.

Aramayo, Roberto R. (1991). "La simbiosis entre ética y filosofía de la historia, o el rostro jánico de la moral kantiana", *Isegoría* 4, 20-36.

Allen, C. & Wallach, W. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

Allen, C. & Wallach, W. (2011). "Moral machines: Contradiction in terms or abdication of human responsibility?" En: Lin, P., Abney, K. & Bekey, G.A. (eds.) *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 55–68). Cambridge, MA: MIT Press.

Apel, Karl-Otto (1985). *La transformación de la filosofía (Tomo II)*. Madrid: Taurus.

Awad, Edmond; Dsouza, Sohan; Kim, Richard; Schulz, Jonathan; Henrich, Joseph; Shariff, Azim; Bonnefon, Jean-François & Rahwan, Iyad (2018). "The Moral Machine experiment". *Nature* (563), 59–64.

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

---

[16] Al respecto, tal y como recogió "The global landscape of AI ethics guidelines" (Jobin, Ienca & Vayena, 2019), la trazabilidad de los principios éticos incluidos en códigos y directrices sobre IA alrededor del mundo arroja 11 principios altamente recurrentes en la orientación de la IA: justicia y bienestar; no-maleficencia, responsabilidad; privacidad; beneficencia; libertad y autonomía; confianza; sostenibilidad; dignidad; solidaridad; y transparencia. No obstante, cabe tener en cuenta que el estudio excluyó los documentos no realizados en lengua inglesa, italiana y griega.

Brown, T. G., Statman, A., & Sui, C. (2021). Public Debate on Facial Recognition Technologies in China. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, https://doi.org/10.21428/2c646de5.37712c5c.

Calvo, Patrici (2023). "Metaverso: aspectos éticos de la tokenización de la economía". *Unisinos Filosofía*, 24(1), 1-20.

Calvo, Patrici (2022). "Gemelos digitales y Democracia". *CLAD. Reforma y Democracia* (82).

Calvo, Patrici (2019). "Etificación, la transformación digital de lo moral", *Kriterion. Revista de Filosofía* (144), 671-688. https://doi.org/10.1590/0100- 512X2019n14409p.

Conill, Jesús (en prensa). Conill, Jesús (2023). "Ética discursiva e Inteligencia artificial. ¿favorece la inteligencia artificial la razón pública?". *Daimon. Revista Internacional de Filosofía* (90), 115-130.

Conill, Jesús (2022). "Razón pública e inteligencia artificial". En Pereira, Gustavo y Pérez Zafrilla, Jesús*, Actualidad de John Rawls en el siglo XXI.* Granada: Comares, pp. 37-47.

Conill, Jesús (2021). *Nietzsche frente a Habermas. Genealogías de la razón.* Madrid: Tecnos.

Conill, Jesús (2019). *Intimidad corporal y persona humana. De Nietzsche a Ortega y Zubiri*. Madrid: Tecnos.

Conill, Jesús (2006). *Ética Hermenéutica. Crítica desde la Facticidad.* Madrid: Tecnos.

Cortina, Adela (1986). *Ética mínima*. Madrid: Tecnos.

Cortina, Adela (1990). *Ética sin moral*. Madrid: Tecnos.

Cortina, Adela (2007). Ética de la razón cordial. Educar en la ciudadanía en el siglo XXI. Oviedo: Nobel.

Cortina, Adela (2010). *Justicia cordial*. Madrid, Trotta.

Cortina, Adela (2017). *Aporofobia, el rechazo al pobre*. Barcelona: Paidós.

Cortina, Adela (2021). *Ética cosmopolita: Una apuesta por la cordura en tiempos de pandemia*. Barcelona, Paidós.

Cortina, A. (2022). "Los desafíos éticos del transhumanismo". *Pensamiento. Revista de Investigación e Información Filosófica*, 78 (298), 471-483. https://doi.org/10.14422/pen.v78.i298. y2022.009.

Cortina, Adela; García-Marzá, Domingo; Conill, Jesús (2003). *Public Reason and Applied Ethics. The Ways of Practical reason in a Pluralist Society*. New York: Routledge.

Cortina, Adela y Martínez, Emilio (1996). *Ética*. Móstoles: Akal.

D'amato, C., Silver, I., Newsome, J., & Latessa, E. (2020). "Progressing policy toward a risk/need informed sanctioning model", *Criminology and Public Policy* (20), 41-69.

Dewitt; Barry; Fischhoff, Baruch & Sahlin, Nils-Eric (2019). "'Moral machine' experiment is no basis for policymaking", *Nature* (567), 31

Donati, Pierpaolo (2019). *Sociología relacional de lo humano*. Barañain: EUNSA.

Feldstein, Steven (2019). *The Global Expansión of AI Surveillance*. Washington, DC: Carnegie.

Fernandes, Daniel L.; Siqueira-Batista, Rodrigo; Gomes, Andréia P.; Souza, Camila R.; da Costa, Israel T.; Cardoso, Felippe da S.L.; Assis, João V. de; Caetano, Gustavo H.L.; Cerqueira, Fabio R. (2017). "Investigation of the visual attention role in clinical bioethics decision-making using machine learning algorithms", *Procedia Computer Science* (108), 1165-1174.

Floridi, Luciano (2012). "Hyperhistory and the Philosophy of Information Policies", *Philosophy & Technology*, *25*, 129-131. https://doi.org/10.1007/s13347-012-0077-4

Floridi, Luciano y Sanders, J.W. (2004). "On the Morality of Artificial Agents", *Minds & Machines* (14), 349-379.

García-Marzá, Domingo (2019). "Ética y democracia. Notas para una renovada ética del discurso" (Ethics and democracy. Notes for renewed discursive ethics). En: González-Esteban, E., Siurana, J.C., López-González, J.L. and García-Granero, M. (eds.), *Ética y Democracia. Desde la razón cordial.* (Ethics and democracy. Based on cordial reason). Granada: Comares, pp. 7-17.

García-Marzá, Domingo (2017). "From ethical codes to ethical auditing: An ethical infrastructure for social responsibility communication", *El profesional de la información, 26*(2), 268-276.

García Marzá, Domingo (1992). Ética de la justicia: J. Habermas y la ética discursiva. Madrid: Tecnos.

García-Marzá, Domingo y Calvo, Patrici (2024). *Algorithmic democracy: A critical perspective based on deliberative democracy*. Cham: Springer.

García-Marzá, Domingo y Calvo, Patrici (2022). "Democracia algorítmica: ¿un nuevo cambio estructural de la opinión pública?". *Isegoría Revista de Filosofía moral y política* (67), 1-15. https://doi.org/10.3989/isegoria.2022.67.17

Habermas, Jürgen (1987). *Theory of Communicative Action Vol. 2 Lifeworld and System: A Critique of Functionalist Reason*. Boston: Beacon Press.

Habermas, Jürgen (1996). *The Inclusion of the Other*. *Studies in Political Theory.* Cambridge, Massachusetts: MIT Press.

Habermas, Jürgen (1984a). *Theory of Communicative Action Vol. 1: Reason and the Rationalization of Society*. Boston: Beacon Press.

Habermas, Jürgen (1984b). *Ciencia y técnica como "ideología"*. Madrid: Tecnos.

Hidalgo César. *Why Information Grows. The Evolution of Order, from Atoms to Economies*. New York, Basic Books. 2015.

High-level expert Group on Artificial Intelligence (2019). *Ethics Guidelines for Trustworthy AI.* Brussels, European Commission. Available at: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

Hooker, John (2018), "Truly Autonomous Machines Are Ethical". *arXiv:1812.02217 [cs.AI]*. https://arxiv.org/pdf/1812.02217.pdf

Jentzsch, S.; Schramowski, P.; Rothkopf, C. A. and Kersting, K. (2019). "Semantics derived automatically from language corpora contain human-like moral choices". En: Conitzer, Vincent (ed.). *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 37–44). New York: AIES, Association for Computing Machinery.

Laniuk, Yevhen (2021). "Social Credit System as a Panopticon: Surveillance and Power in the Digital Age". En: Denys Kiryukhin (Ed.), *Community and Tradition in Global Times* (pp. 211-234). Washington D.C: Cultural Heritage and Contemporary Change.

Johnston, Lachlan (2018). "There's an AI Running for the Mayoral Role of Tama City, Tokyo", *Otaquest*, http://www.otaquest.com/tama-city-ai-mayor/.

Kant, Inmanuel (1968a). *Kritik der reinen Vernunft* (1. Auflage 1781), Kants Werke. Berlin: Akademie Textausgabe, Walter de Gruyter, IV, 1-252

Kant, Inmanuel (1968b). *Kritik der reinen Vernunft* (2. Auflage 1787), Kants Werke. Berlin: Akademie Textausgabe, Walter de Gruyter, IV, 1-252. https://doi.org/10.3390/su12072824

López-González, J.L. (2022). *La ética ante la cinética del turismo. Aportaciones desde la teoría crítica de la resonancia de Harmunt Rosa.* Castellón, Universitat Jaume I.

Matsumoto, Tetsuzo (2018). *The Day AI Becomes God. The Singularity will Save Humanity*. Cambridge (NZ): Media Tectonics.

Mayer-Schöenberger, Viktor & Cukier, Kenneth N. (2013). *Big Data. A Revolution that Will Transform How We Live, Work, and Think*. Londres: John Murray Publishers.

Motta, Luís Claudio de Souza; Oliveira, Lucas Nicolau de; Silva, Eugenio; Siqueira-Batista, Rodrigo (2016). "Toma de decisiones en (bio)ética clínica: enfoques contemporáneos", *Revista de Bioética 24*(2), 304-331.

Pérez-Zafrilla. Jesús (2021). "Polarización artificial: cómo los discursos expresivos inflaman la percepción de polarización política en internet", *Recerca. Revista de Pensament i Anàlisi* 26 (2), 1-23.

Rapaport, W. (2005). "Philosophy of Computer Science: An Introductory Cours". *Teaching Philosophy* 28 (4).

Saura, C. (2022). "El lado oscuro de las GAFAM: monopolización de los datos y pérdida de privacidad". *Veritas. Revista de Filosofía y Teología* (52), 28-46.

Schalkoff, R. J. (1990). Artificial Intelligence: An Engineering Approach. Michigan: McGraw-Hill.

Searle, John (1988). *Mind, Brain, and Behaviour.* New York, Routledge.

Siqueira-Batista, Rodrigo; Gomes, Andréia Patrícia; Mendes Maia, Polyana; Teoldo da Costa, Israel; Oliveira de Paiva, Alcione, Ribeiro Cerqueira, Fábio (2014). "Los modelos de toma de decisiones en bioética clínica: apuntes para un enfoque computacional", *Revista de Bioética 22*(3), 456-461.

Tigard, D. (2021). "Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible". *Cambridge Quarterly of Healthcare Ethics 30*(3), 435-447.

van Dijck, José (2014), "Datafication, Dataism and Dataveillance: Big Data between Scientific Paradigm and Ideology", *Surveillance and Society 12*(2), 197-208.

Zuboff, Shoshana (2019). *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power.* New York: PublicAffairs.