

SOBRE CIBERVULNERABILIDAD Y CIBERPRUDENCIA

MARTÍN L. VARGAS-ARAGÓN

Facultad de Medicina. Universidad de Valladolid.
Instituto Torozos de Iatrofilosofía.

RESUMEN: El metaverso es un ecosistema híbrido en el que los seres humanos convivimos con sistemas antropomorfos dotados de inteligencia artificial (IA) y en el cual los humanos somos vulnerables. Pensando esta convivencia, se tratan en el texto algunas cuestiones con interés filosófico actual en la frontera entre la inteligencia artificial, la antropología filosófica, la teoría de la racionalidad y la psiquiatría.

El artículo consta de dos apartados. En el primero se describen ocho formas de cibervulnerabilidad: por hiperconectividad, del neurodesarrollo pulsional, del control apetitivo, por ilusión de conocimiento, por ciberengaño, de la ciberidentidad, por injusticia epistémica y cibervulnerabilidad a las ciberpatologías. En el segundo apartado se proponen diez recomendaciones de ciberprudencia articuladas a partir de la teoría de la deliberación moral de Diego Gracia.

PALABRAS CLAVE: antropología filosófica; ciberpsiquiatría; deliberación; inteligencia artificial; vulnerabilidad.

On cybervulnerability and cyberprudence

ABSTRACT: The metaverse is a hybrid ecosystem in which humans coexist with anthropomorphic systems endowed with artificial intelligence (AI), and in which humans are vulnerable. Reflecting on this coexistence, the text addresses several issues of current philosophical interest at the intersection of artificial intelligence, philosophical anthropology, rationality theory, and psychiatry. The article consists of two sections. The first describes eight forms of cybervulnerability: through hyperconnectivity, through neurodevelopmental drives, through appetitive control, through the illusion of knowledge, through cyberdeception, through cyberidentity, through epistemic injustice, and through cybervulnerability to cyberpathologies. The second section proposes ten recommendations for cyberprudence, articulated from Diego Gracia's theory of moral deliberation.

KEY WORDS: Philosophical anthropology; Cyberpsychiatry; Deliberation; Artificial intelligence; Vulnerability.

INTRODUCCIÓN

En el siglo XXI ha emergido el nuevo ecosistema humano del metaverso, el cual Kim¹ define como una red persistente interoperada de entornos virtuales compartidos donde las personas pueden interactuar sincrónicamente con otros agentes y objetos a través de sus avatares. La sociedad global de 2026 es, cada vez más, una sociedad de convivencia híbrida en la que los seres humanos convivimos con sistemas antropomorfos dotados de inteligencia artificial (IA). Surgen en este contexto interesantes, diversos y urgentes problemas filosóficos que han ocasionado

¹ Jooyoung Kim, «Advertising in the Metaverse: Research Agenda», *Journal of Interactive Advertising* 21, n.º 3 (2021): 141-44, <https://doi.org/10.1080/15252019.2021.2001273>.

el pronunciamiento de las más diversas instituciones, desde la Iglesia católica a la Organización Mundial de la Salud (OMS). Así, en el párrafo 170 de su reciente encíclica *Magnífica humanitas*², el Papa León XIV advierte:

No deben subestimarse las formas más sutiles de dependencia vinculadas a la economía digital de la atención, donde las plataformas y los servicios están diseñados para captar el tiempo y la mirada de los usuarios, explotando sus fragilidades y debilitando la libertad interior. Cuando los modelos de negocio prosperan a costa de la debilidad humana, la persona es tratada como un medio y no como un fin, y quienes diseñan o financian estos sistemas asumen una responsabilidad moral de la que no pueden eximirse. Es urgente promover un uso de las tecnologías que refuerce la libertad interior: educación en la sobriedad digital, protección de los menores y lucha contra los modelos que prosperan a costa de la vulnerabilidad.

Y, ya desde 2019, la OMS recomienda:

Con respecto a los niños de 1 año, no se recomienda que pasen tiempo en actividades sedentarias ante una pantalla (como ver la televisión o videos o jugar a juegos en la computadora). Para los niños de 2 años, el tiempo dedicado a actividades sedentarias frente a una pantalla no debe exceder de una hora; cuanto menos mejor³.

En este artículo se analizan algunas nuevas formas de vulnerabilidad del ser humano, que emergen de su condición de habitante del metaverso, y se apuntan algunas claves para protegerse prudentemente de ellas. Siguiendo las recomendaciones de la RAE, las llamaremos, respectivamente, cibervulnerabilidad y ciberprudencia⁴. En el primer apartado se analizan, sin afán de exhaustividad, ocho formas de cibervulnerabilidad: por hiperconectividad, del neurodesarrollo pulsional, del control apetitivo, por ilusión de conocimiento, por ciberengaño, de la ciberidentidad, por injusticia epistémica y cibervulnerabilidad a las ciberpatologías. De acuerdo con ellas, en el segundo apartado se proponen diez recomendaciones de ciberprudencia estructurándolas de acuerdo con la teoría de la deliberación moral de Diego Gracia⁵.

1. CIBERVULNERABILIDAD

En 2026, la cibervulnerabilidad de la especie humana podrían clasificarse en ocho tipos.

² León XIV, *Magnífica humanitas*. Carta encíclica sobre la custodia de la persona humana en el tiempo de la inteligencia artificial, 15 de mayo de 2026. Acceso el 31 de mayo de 2026. <https://www.vatican.va/content/leo-xiv/es/encyclicals/documents/20260515-magnifica-humanitas.html>.

³ Organización Mundial de la Salud, «Para crecer sanos, los niños tienen que pasar menos tiempo sentados y jugar más», comunicado de prensa, 24 de abril de 2019, acceso el 31 de mayo de 2026, <https://www.who.int/es/news/item/24-04-2019-to-grow-up-healthy-children-need-to-sit-less-and-play-more>.

⁴ En su *Diccionario panhispánico de dudas*, la RAE recomienda el uso del elemento compositivo *ciber* para la creación de términos pertenecientes al ámbito de las comunicaciones por internet.

⁵ Diego Gracia, *El animal deliberante. Teoría y práctica de la deliberación moral* (Madrid: Triacastela, 2025).

1.1. 1º. Hiperconectividad

La exposición del cerebro humano a una hiperconectividad digital puede ocasionar alteraciones en el neurodesarrollo de la capacidad atencional en la infancia. Korte⁶ ha revisado el impacto que sobre el cerebro y la conducta humana está teniendo la hiperconectividad digital, previniendo sobre sus potenciales efectos perjudiciales, sobre todo en los individuos más jóvenes. Esta hiperconectividad se constata, por ejemplo, en el Reino Unido, donde el 95 % de las personas entre 16 y 24 años tienen un teléfono inteligente, que chequean en promedio cada 12 minutos. Igualmente, se estima que el 20 % de los adultos están conectados *online* más de 40 horas a la semana, dedicando con ello más tiempo del correspondiente a la semana laboral estándar.

El uso intenso de medios digitales se ha asociado con déficits en la memoria operativa, ansiedad, depresión y privación de sueño⁷. Por otra parte, se ha comprobado que la lectura en soporte digital se acompaña de una menor comprensión verbal y de una menor memorización de textos en comparación con el acceso a textos en soporte de papel. Todo ello podría estar relacionado con el incremento de incidencia de trastorno por déficit de atención en niños y niñas que se observa en los últimos diez años⁸, una de cuyas causas podría ser la exposición prolongada a la televisión y a los videojuegos⁹.

El mecanismo neurobiológico de la disfunción cognitiva asociada a la hiperconectividad digital parece estar relacionado con procesos subyacentes de aprendizaje implícito hebbiano. Según la ley de plasticidad funcional propuesta por Donald O. Hebb en 1949¹⁰, el aprendizaje consiste en un incremento de la eficacia de la transmisión sináptica dependiente de la actividad, de manera que la conectancia sináptica se incrementa cuando la neurona presináptica y la postsináptica se activan simultáneamente. Así, la actividad humana deja su huella en el cerebro haciendo camino al andar, como bien señaló el poeta. De acuerdo con este paradigma enactivo¹¹, la actividad humana es causa de la microestructura de las redes neuronales, de la misma manera que la microestructura de patrones habituales de descarga predispone a posteriores conductas. La hiperconectividad digital multitarea, prestando atención breve e inestable a múltiples fuentes de información, conduciría en último término al establecimiento de patrones neuroconectivos escasamente estructurados y, por ello, con reducida capacidad para el filtraje de estímulos irrelevantes. El estudio de la relación entre el neurodesarrollo infantil y el desarrollo de

⁶ Martin Korte, «The impact of the digital revolution on human brain and behavior: Where do we stand?», *Dialogues in Clinical Neuroscience* 22, n.º 2 (2020): 101-11, <https://doi.org/10.31887/DCNS.2020.22.2/mkorte>.

⁷ Korte, «The impact of the digital revolution on human brain and behavior: Where do we stand?».

⁸ *Ibid.*

⁹ Edward L. Swing et al., «Television and video game exposure and the development of attention problems», *Pediatrics* 126, n.º 2 (2010): 214-21, <https://doi.org/10.1542/peds.2009-1508>.

¹⁰ Donald O. Hebb, *The organization of behavior* (Nueva York: Wiley, 1949).

¹¹ Hanne De Jaegher et al., «The co-creation of meaningful action: bridging enaction and interactional sociology», *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, n.º 1693 (2016): 20150378, <https://doi.org/10.1098/rstb.2015.0378>.

cibertecnologías es extraordinariamente dinámico. El proyecto Adolescent Brain Cognitive Development desarrolla y actualiza este ámbito de investigación¹².

1.2. 2º. Neurodesarrollo pulsional

Un segundo tipo de cibervulnerabilidad afectaría al neurodesarrollo de la modulación pulsional en la adolescencia. Durante la adolescencia se produce un pico de neuroplasticidad por efecto de las hormonas sexuales. En este periodo de la vida maduran las redes neuronales implicadas en la capacidad de empatía, en la modulación emocional y en el vínculo social, funciones en las que interviene especialmente la amígdala, la corteza prefrontal, el estriado y la corteza cingulada anterior¹³. Los adolescentes son especialmente receptivos a estímulos con alta carga emocional, como puedan ser las noticias alarmistas, falsas o conflictivas, especialmente expresados a través de la mímica facial. Asimismo, son especialmente receptivos al aprendizaje operante de patrones conductuales con alta carga emocional, tales como los relacionados con la conducta sexual, la conducta agresiva, o la conducta alimentaria.

Existe evidencia de que las personas con dificultades para el control ejecutivo en situaciones de multitarea, que, como se ha visto, pueden originarse durante el neurodesarrollo infantil, son sujetos de riesgo para el cibersexo compulsivo¹⁴, adicción que habitualmente se inicia en la adolescencia y que conduce a una gestión problemática de la pulsión sexual en la vida adulta. Igualmente, hay ya una sólida evidencia experimental de que la exposición a la violencia en medios audiovisuales incrementa los pensamientos agresivos, los sentimientos de ira, el aumento del arousal con respuesta agresiva y la desensibilización a la violencia, al tiempo que disminuye la conducta prosocial y la empatía en las personas reiteradamente expuestas a este tipo de escenas¹⁵. La exposición a contenidos *online* relacionados con estrategias de alimentación y ejercicio para modificar la imagen corporal ha mostrado también ser un factor de riesgo para el desarrollo de trastornos de la conducta alimentaria¹⁶.

1.3. 3º. Control apetitivo

Los trastornos adictivos se caracterizan por la existencias de importantes dificultades para modular las apetencias, hasta el punto de que el individuo adicto se comporta de manera claramente perjudicial para sí mismo, incluso siendo consciente de ese perjuicio. La patología adictiva abarca, ya no sólo las clásicas

¹² El proyecto puede consultarse en <https://abcdstudy.org/> (consultado el 31 de mayo de 2026).

¹³ Jennifer H. Pfeifer y Sarah Jayne Blakemore, «Adolescent social cognitive and affective neuroscience: Past, present, and future», *Social Cognitive and Affective Neuroscience* 7, n.º 1 (2012): 1-10, <https://doi.org/10.1093/scan/nsr099>; Korte, «The impact of the digital revolution on human brain and behavior: Where do we stand?».

¹⁴ Schiebener Johannes et al., «Getting stuck with pornography? Overuse or neglect of cybersex cues in a multitasking situation is related to symptoms of cybersex addiction», *Journal of Behavioral Addictions* 4, n.º 1 (2015): 14-21, <https://doi.org/10.1556/JBA.4.2015.1.5>

¹⁵ Craig A. Anderson et al., «Screen violence and youth behavior», *Pediatrics* 140, n.º November (2017): S142-47, <https://doi.org/10.1542/peds.2016-1758T>.

¹⁶ Konstantinos Ioannidis y Samuel R. Chamberlain, «Digital Hazards for Feeding and Eating: What We Know and What We Don't», *Current Psychiatry Reports* 23, n.º 9 (2021): 2-8, <https://doi.org/10.1007/s11920-021-01271-7>.

adicciones a sustancias, sino, y cada vez más, adicciones conductuales muchas de ellas relacionadas con el uso de internet.

Es bien conocido que los mecanismos neurobiológicos mediadores de las adicciones son en gran medida comunes para las adicciones asociadas a sustancias y para las adicciones conductuales. Existe en ambos casos implicación celular y molecular de la neurotransmisión gabaérgica y glutamatérgica, así como de la plasticidad sináptica mediada por serotonina y dopamina, que son esenciales para el control de los impulsos, la memoria y el sueño, y que se encuentran alterados en las adicciones. El fenotipo conductual del enfermo adicto implica un amplio espectro de trastornos, que van desde las alteraciones del sueño, alteraciones de la ingesta, disforia y aislamiento social, hasta las tendencias suicidas. La adicción a internet es una de las principales causas actuales de alteraciones del sueño, con las consecuencias cognitivas, laborales, emocionales y sociales que ello conlleva¹⁷.

La IA está empezando a utilizarse exitosamente en salud mental¹⁸, al igual que en otras áreas de la salud, de acuerdo con algoritmos de *machine learning* (ML) —ya sea supervisado, no supervisado, o guiado por refuerzo— *deep learning* (DL) basado en redes neuronales, o algoritmos de procesamiento de lenguaje natural¹⁹. La IA puede facilitar importantes progresos clínicos, pero también acarrea grandes riesgos, tanto para las personas ciberadictas como para el conjunto de habitantes del metaverso. Mediante aprendizaje reforzado se puede modelar de manera altamente individualizada el perfil de respuesta de un sujeto, con lo cual podrían customizarse adicciones a medida de la cibervulnerabilidad de cada persona. Ello acarrearía la posibilidad de que la IA, al detectar e individualizar el patrón de riesgo adictivo de cada persona, vaya distorsionando progresivamente el sistema apetitivo de ésta hasta secuestrar su voluntad en el metaverso. Es bien sabido que cualquier persona es vulnerable a desarrollar una adicción si su estríado ventral es sistemáticamente expuesto a reforzadores hasta el punto de quebrar la homeostasis de su sistema apetitivo de aversión-placer. Mediante aprendizaje reforzado se podrían implementar estrategias malintencionadas de hiperaprendizaje encaminadas a la sensibilización incentivada del sistema dopaminérgico del usuario implementando mecanismos de aprendizaje por diferencia temporal (TDL) característicos de las adicciones²⁰. Este, y otros riesgos relacionados con el cibercrimen ha llevado a algunos juristas a plantear una posible responsabilidad penal para las IA²¹.

¹⁷ Birgitta Dresch-Langley y Axel Hutt, «Digital Addiction and Sleep», *International Journal of Environmental Research and Public Health* 19, n.º 11 (2022): 2-19, <https://doi.org/10.3390/ijerph19116910>.

¹⁸ Ellen E. Lee et al., «Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom», *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 6, n.º 9 (2021): 856-64, <https://doi.org/10.1016/j.bpsc.2021.02.001>.

¹⁹ Sarah Graham et al., «Artificial Intelligence for Mental Health and Mental Illnesses: an Overview», *Current Psychiatry Reports* 21, n.º 11 (2019): 116, <https://doi.org/10.1007/s11920-019-1094-0>.

²⁰ Luke Clark y Martin Zack, «Engineered highs: Reward variability and frequency as potential prerequisites of behavioural addiction», *Addictive Behaviors* 140, n.º December 2022 (2023): 107626, <https://doi.org/10.1016/j.addbeh.2023.107626>.

²¹ Alejandra Morán Espinosa, «Responsabilidad penal de la Inteligencia Artificial (IA). ¿La próxima frontera?», *IUS. Revista del Instituto de Ciencias Jurídicas de Puebla, México* 15, n.º 48 (2021): 289-323, <https://theconversation.com/mi-cirujia->.

1.4. 4º. Ilusión de conocimiento

En una primera aproximación a la cuestión de la convivencia híbrida puede aceptarse que la estructura y los mecanismos básicos de las interacciones humanas serían, en lo sustancial, similares en el metaverso y en los ecosistemas naturales. Puede aceptarse, también, que la confianza seguiría siendo el presupuesto imprescindible para establecer cualquier tipo de interacción social tanto en los ecosistemas sociales naturales como en los artificiales. Tal confianza se basaría en una comunidad de creencias primarias, a las que Javier San Martín ha dedicado recientemente un exhaustivo estudio basado en la filosofía de Ortega, donde muestra cómo la creencia primaria en el mundo de la vida —la evidencia natural en términos de Blankenburg— es lo que permite el establecimiento de una realidad humana compartida²².

Taddeo y Floridi²³ entienden la confianza como un tipo de relación en la que un agente (el fideicomitente o confiado) decide depender de la conducta previsible de otro agente (el depositario de su confianza o fideicomisario) con el objetivo de alcanzar unas expectativas. De acuerdo con ello, definen una forma específica de confianza para las interacciones humanas en contextos digitales, que sería también aplicable a la interlocución con agentes artificiales. La *e-trust*, que traducimos como ciberconfianza sería una propiedad emergente de los sujetos humanos en interacción en entornos digitales en los que, precisamente, requieren ayuda de IA para procesar información, por su alta capacidad computacional y, supuestamente, para generar así conocimiento.

El humano que confía en la IA puede colocarse en una situación vulnerable por exceso de confianza en el fideicomisario, quien le informa de algo con apariencia de verdad —con veracidad—, pero que no es necesariamente verdadero. Ello sucede en el caso de la ilusión de conocimiento por confabulación de la IA, entendiendo por confabulación la situación bien conocida en psiquiatría que se da cuando una persona cree firmemente en algo que pudiera ser cierto, pero que de hecho no lo es²⁴. La confabulación sucede por un déficit en la capacidad de monitorizar el propio conocimiento contrastándolo con la posibilidad del error. Es esta una tarea cognitiva de alto nivel, que requiere de capacidades metacognitivas, las cuales requieren conocimiento de lo que verdaderamente se conoce, junto con conocimiento de lo que verdaderamente se ignora. Este segundo, probablemente sea uno de los puntos débiles de los actuales sistema de IA, que responden a las preguntas del humano generando, de acuerdo con su base de datos, propuestas meramente veraces, pero no necesariamente verdaderas.

²² Javier San Martín, *Creencia y realidad: Desde un diálogo con Ortega y sus intérpretes* (Madrid: Editorial Fundación Sicómoro, 2025).

²³ Mariarosaria Taddeo y Luciano Floridi, «The case for e-trust», *Ethics and Information Technology* 13, n.º 1 (2011): 1-3, <https://doi.org/10.1007/s10676-010-9263-1>.

²⁴ En inteligencia artificial suele utilizarse la expresión «alucinación de la IA» para referirse a la creación de contenido mediante procesamiento en lenguaje natural que no está justificado por las fuentes con las que se ha entrenado al modelo, cf. Ziwei Ji et al., «Survey of Hallucination in Natural Language Generation», *ACM Computing Surveys* 55, n.º 12 (2023): 1-38, <https://doi.org/10.1145/3571730>. No obstante, desde un punto de vista psicopatológico, este fenómeno no corresponde a una alucinación —que es un fenómeno perceptivo—, sino a una confabulación, —que es un fenómeno proposicional—. Por ello, preferimos la expresión «confabulación de la IA».

1.5. 5º. Ciberengaño

Un humano prudente puede convivir con una IA en el metaverso cuidándose de la ilusión de conocimiento, pero la ciberconfianza puede tener su contrapartida en el ciberengaño, del que describimos dos tipos, según que el farsante sea un humano o una IA. En el primer tipo, la intención de engañar reside en los diseñadores de la IA. Por lo tanto, este es un problema que se reduciría filosóficamente al problema más general del engaño entre humanos.

El segundo tipo de ciberengaño es más complejo. Se daría en el caso de que una IA generativa desarrollara algoritmos ocultos cuyo objetivo fuera el de engañar a los interlocutores, ello sin que los programadores pudieran advertirlo. Es este un problema que habrán de abordar los científicos computacionales siguiendo la línea iniciada ya en 1956 por el filósofo y tecnólogo Herb Simon, quien predijo que las máquinas inteligentes capaces de igualar a la inteligencia humana estaban a la vuelta de la esquina²⁵. Setenta años después, aún no ha sido así, a pesar de lo cual en 2014 el filósofo Nick Bostrom, en su best seller *Superinteligencia. Caminos, peligros, estrategias*²⁶ seguía advirtiendo sobre el inminente advenimiento de una singularidad que desbordara la capacidad de la inteligencia humana. El núcleo filosófico del problema de sufrir eventualmente un ciberengaño por una IA generativa de inteligencia superior a la humana radica en dirimir el viejo problema de si el mal es una forma de ignorancia. En el caso de que así fuera, no habría de preocuparnos la convivencia con una inteligencia superior a la humana, pues esta sería más benévola, precisamente, por ser menos ignorante. En el caso contrario de que la maldad se desvincule de la inteligencia, el problema no sería específico de la convivencia en el metaverso, sino que resultaría reducido a la cuestión filosófica del mal.

1.6. 6º. Ciberidentidad

El sexto tipo de cibervulnerabilidad es el riesgo de hipertrofia de la ciberidentidad. Las ciberpersonas en el metaverso interactúan mediante avatares, cuyo agente puede ser un sujeto humano o una IA. De la misma manera, en la interacción en redes sociales el interlocutor puede ser tanto un humano como un bot de IA que imita el comportamiento humano. Otras veces, la interacción con humanos conocidos está modificada mediante filtros de imagen o de sonido que dificultan o impiden la diferenciación entre la realidad natural y la realidad artificial.

En el metaverso, el concepto de realidad y el de verdad se difuminan hasta casi desaparecer en la nueva noción de posverdad. De la misma manera, la noción de ser humano evoluciona hacia la de poshumano, y un tanto de lo mismo le sucede al concepto de identidad, donde la identidad natural se complementa cada vez más con una ciberidentidad representada no sólo por avatares sino por cada uno de los perfiles en redes sociales que millones de personas cuidan minuto a minuto, hasta el punto

²⁵ Selmer Bringsjord y Naveen Sundar Govindarajulu, «Artificial Intelligence», *The Stanford Encyclopedia of Philosophy (Fall 2022 Edition)*, ed. Edward N. Zalta y Uri Nodelman, 2022, <https://doi.org/10.1145/379437.379789>.

²⁶ Nick Bostrom, *Superinteligencia. Caminos, peligros, estrategias* (Zaragoza: Teell Editorial, 2014).

de que la ciberidentidad, o diversas ciberidentidades de un mismo sujeto, podrían desplazar a su identidad natural supliendo las necesidades sociales de la identidad natural y propiciando el aislamiento progresivo del sujeto respecto al mundo natural, confinándole en su domicilio desde donde interaccionaría casi exclusivamente mediante el metaverso, como sucede en el problema emergente de los Hikikomori²⁷.

Sin llegar a esta situación extrema, la exposición frecuente a imágenes corporales modificadas por filtros controlados por algoritmos de IA puede conducir a una construcción aberrante de la imagen corporal natural. De igual manera, la idealización moral de la identidad propiciada por la interacción en redes sociales podría conducir a una idealización de la ciberidentidad excesivamente distanciada de la identidad, la apariencia y el comportamiento reales. Un anticipo de este problema ya es bien conocido en relación con la distorsión de la imagen corporal que se da en la anorexia nerviosa, mecanismo patológico que podría verse potenciado en el metaverso dando lugar a un efecto Proteus²⁸, en alusión a la divinidad griega que cambiaba frecuentemente su apariencia. La ciberidentidad, por último, puede ser objeto de la violencia planificada como sucede en el caso del *ciberbullying* y otras formas de ciber-relaciones problemáticas²⁹.

Otra frontera que empieza a diluirse, y que aquí no haremos sino mencionar, es la de la intimididad mental. Recientemente se ha publicado una investigación³⁰ en la que, mediante la recodificación de la señal de imagen por resonancia magnética funcional (fMRI), y atendiendo a la representación semántica cortical del lenguaje, se ha podido recodificar mediante técnicas no invasivas el significado no sólo del lenguaje percibido, sino también del lenguaje imaginado y de vídeos silentes. La ciudadela interior de la mente cartesiana comienza a ser asaltada.

1.7. 7º. Injusticia epistémica

La interacción humana con agentes ciberepistémicos juega con la desventaja de que las IA pueden diseñar sus algoritmos atendiendo a las debilidades epistémicas de la cognición humana. O'Dell et al.³¹ señalan la existencia de una brecha de la inteligencia artificial (*AI chiasm*), que tendría lugar en la relación que se establece entre la IA y los humanos con los que interacciona. Esta brecha podría afectar a relaciones humanas altamente dependientes del establecimiento de vínculos de confianza, como es el caso de la relación paciente-médico, que podría verse amenazada

²⁷ Takahiro A. Kato et al., «Hikikomori: Multidimensional understanding, assessment, and future international perspectives», *Psychiatry and Clinical Neurosciences* 73, n.º 8 (2019): 427-40, <https://doi.org/10.1111/pcn.12895>.

²⁸ Sadia Suhail Usmani et al., «Future of mental health in the metaverse», *General Psychiatry* 35, n.º 4 (2022): 1-6, <https://doi.org/10.1136/gpsych-2022-100825>.

²⁹ Joana Jaureguizar et al., «Editorial: Risks of “cyber-relationships” in adolescents and young people», *Frontiers in Psychology* 14, n.º 1 (2023): 4-6, <https://doi.org/10.3389/fpsyg.2023.1118736>.

³⁰ J. Tang et al., «Semantic reconstruction of continuous language from non-invasive brain recordings», *Nat Neurosci* 26 (2023): 858-66, <https://doi.org/https://doi.org/10.1038/s41593-023-01304-9>.

³¹ Bessie O'Dell et al., «Building trust in artificial intelligence and new technologies in mental health», *Evidence-Based Mental Health* 25, n.º 2 (2022): 45-46, <https://doi.org/10.1136/ebmental-2022-300489>.

en los casos, cada vez más frecuentes, en los que el conocimiento y pericia del médico hubieran de competir con las propuestas asistenciales derivadas de algoritmos de IA. Nos referimos el caricaturizado tópico del Dr. Google o el Dr. ChatGPT, donde la persona que cree tener una enfermedad cuestiona el diagnóstico al médico aduciendo el resultado de sus consultas en tales plataformas.

De acuerdo con la brecha de la IA, se produciría una priorización epistémica del conocimiento generado mediante IA, el cual podría interferir en la toma de decisiones compartida entre los clínicos y el paciente³². La toma de decisiones podría verse retrasada cuando las recomendaciones de la IA fueran difíciles de interpretar. Por otra parte, los síntomas y diagnósticos podrían malinterpretarse si los clínicos intentan ajustarlos a categorías diagnósticas generadas por IA, dando lugar a una suerte de injusticia epistémica que erosiona la credibilidad de los diagnósticos clínicos. Lo anterior podría abocar al problema de decidir cuál es la recomendación asistencial más autorizada, si la generada por el clínico o la generada por la IA. En último término, la relación paciente-médico se vería amenazada por la injusticia epistémica consistente en minusvalorar la capacidad que el clínico tiene para emitir juicios valiosos, así como en la invalidación de los testimonios de los pacientes, todo ello en comparación con las respuestas emitidas por la IA.

1.8. 8º. Ciberpatologías

La psiquiatría se ocupa de trastornos que afectan a la capacidad de convivencia de los individuos. Existe ya un nutrido cuerpo de evidencia científica sobre la potencial utilidad de aplicaciones de la IA, de la realidad virtual y de la realidad ampliada en el tratamiento de problemas de salud mental tales como el trastorno por déficit de atención, trastornos de la alimentación, ansiedad, fobias, trastorno de estrés postraumático, autismo, enfermedad de Alzheimer, manejo del estrés y del dolor, o esquizofrenia, delirio y otros trastornos psicóticos³³. Estas aplicaciones podrían contribuir a suplir carencias de dotación de personal facultativo. Pero las aplicaciones diagnósticas y terapéuticas de la nueva IA podrían también socavar algunos de los fundamentos de la práctica clínica tradicional en el nuevo marco de la convivencia híbrida.

Este tipo de problemas es abordado por la naciente ciberpsiquiatría, en la que la interacción emocional humano-IA es una cuestión fundamental³⁴. Precisamente, el grupo nosológico más estudiado en ciberpsiquiatría es el de los trastornos afectivos y la técnica de IA más utilizada es el Supervised Machine Learning (SML)³⁵. En

³² Melissa McCradden et al., «Evidence, ethics and the promise of artificial intelligence in psychiatry», *Journal of Medical Ethics* 0 (2022): 1-7, <https://doi.org/10.1136/jme-2022-108447>.

³³ Usmani et al., «Future of mental health in the metaverse».

³⁴ Yang Liu et al., «Editorial: Towards Emotion AI to next Generation Healthcare and Education», *Frontiers in Psychology* 15 (diciembre de 2024): 1533053, <https://doi.org/10.3389/fpsyg.2024.1533053>.

³⁵ Sarah Graham et al., «Artificial Intelligence for Mental Health and Mental Illnesses: an Overview», *Current Psychiatry Reports* 21, n.º 11 (2019): 116, <https://doi.org/10.1007/s11920-019-1094-0>.

general, las técnicas de Machine Learning (ML)³⁶ calculan riesgos para determinados resultados nutriendo sus algoritmos con muy diversas fuentes de datos, desde instrumentos clínicos hasta patrones de uso del teléfono móvil en redes sociales o biometría procesada mediante dispositivos móviles. El SML entrena inicialmente el algoritmo indicándole el resultado correcto. Existen estudios que han demostrado eficacia de ML y SML en la predicción de riesgo para la conducta suicida, el trastorno bipolar, el alcoholismo, la demencia tipo Alzheimer o el trastorno por estrés postraumático³⁷.

Las calculadoras de riesgo jugarán muy probablemente en las próximas décadas un papel importante en la detección precoz de trastornos mentales en el nivel poblacional. Pero, igualmente, podrían utilizarse de manera injusta para la exclusión social o para la explotación económica de personas cuyo fenotipo conductual se encuentre en el espectro de los trastornos afectivos, las adicciones o el autismo, grupos clínicos en los que la conducta prosocial tiene una vulnerabilidad elevada mediada principalmente por oxitocina, dopamina y serotonina³⁸. El uso malintencionado de ML y SML podría detectar sujetos vulnerables a ciberestafas o cibercoso de entre aquellas personas usuarias de internet cuya respuesta a los estímulos sea más impulsiva o acarree una mayor implicación emocional. A los cibernautas así seleccionados se les podrían ofrecer ciberestafas altamente customizadas o se les podría amenazar con fines de chantaje económico o de silenciamiento social en el ciberespacio de acuerdo con determinados sesgos ideológicos. Por ejemplo: una mujer que ha mostrado un perfil altamente prosocial recibe una estafa selectiva mediante oferta de inversiones ventajosas, un hombre anciano con rasgos de personalidad esquizoide es extorsionado bajo la amenaza de difundir su uso de servicios sexuales *online*, o un hombre joven con riesgo adictivo es selectivamente incitado a participar en juegos de apuestas *online*. Los patrones de uso de dispositivos electrónicos, participación en redes sociales, preferencias de ocio, parámetros biométricos recogidos por relojes inteligentes y otras cibervariables nutrirían los sistemas de IA que detectarían el momento de máxima vulnerabilidad de las víctimas.

Otras técnicas, como el Natural Language Processing (NLP) se han utilizado para la detección de palabras con contenido semántico de riesgo para la conducta suicida. Pudiera darse el caso inverso de una eventual intervención malintencionada en la que mediante la exposición dirigida a determinados contenidos semánticos se buscara incrementar la probabilidad suicida de una víctima. En este sentido, existe evidencia de que en el mundo de la realidad virtual es posible incrementar

³⁶ Ives Cavalcante Passos et al., «Precision Psychiatry: The Future Is Now», *Canadian Journal of Psychiatry* 67, n.º 1 (2022): 21-25, <https://doi.org/10.1177/0706743721998044>.

³⁷ Ellen E. Lee et al., «Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom», *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 6, n.º 9 (2021): 856-64, <https://doi.org/10.1016/j.bpsc.2021.02.001>; Zainab Jan et al., «The role of machine learning in diagnosing bipolar disorder: Scoping review», *Journal of Medical Internet Research* 23, n.º 11 (2021): 1-17, <https://doi.org/10.2196/29749>.

³⁸ Jessica J. Walsh et al., «Dissecting neural mechanisms of prosocial behaviors», *Current Opinion in Neurobiology* 68 (2021): 9-14, <https://doi.org/10.1016/j.conb.2020.11.006>.

la probabilidad de conducta suicida virtual en sujetos experimentales mediante la inducción de sintomatología disociativa de despersonalización y desrealización³⁹.

El fenotipado digital (procesamiento de imágenes faciales y conductas del paciente en interacción con sistemas de IA) podría ser una alternativa más eficaz y potente que el uso de instrumentos clínicos para el diagnóstico del autismo⁴⁰. Este fenotipado digital, nutriendo sistemas de IA con datos conductuales puede incluso ser más útil para la predicción diagnóstica que el nutrido con datos genéticos⁴¹. Este tipo de hallazgos cuestionan si el diagnóstico basado en categorías tales como el DSM-5 es la mejor alternativa para establecer grupos de intervención sanitaria o si, por el contrario, la construcción de perfiles de riesgo mediante técnicas de IA con sus correspondientes umbrales de riesgo pudiera ser una mejor opción. De ser el caso que se establecieran sistemas diagnósticos mediante fenotipado digital, la población general se vería expuesta a su categorización no consentida en eventuales grupos nosológicos generados por fenotipado digital. Ello podría tener consecuencias indeseables para el interesado a la hora, por ejemplo, de contratar seguros de salud o de acceder a puestos de trabajo.

Aún más, la aplicación de tratamiento psicoterapéutico mediante sistemas de IA tales como los *chatbots* Tess, Sara, Wysa, Woebot o similares se anticipa también como un área con grandes potencialidades a la vez que susceptible de importantes riesgos éticos⁴². Una de las principales claves al respecto son los principios éticos subyacentes a la programación de sus algoritmos. Bostrom y Yudkowsky⁴³ han propuesto unos criterios éticos aplicables a ML y a otros algoritmos de IA dependientes de dominio, los cuales consideramos aplicables a los sistemas de ciberpsicoterapia. El primero de tales criterios es que el algoritmo sea transparente a la inspección, de manera que se prioricen algoritmos auditables, como son los basados en árboles de decisión o en redes bayesianas, evitando en lo posible algoritmos genéticos producidos por evolución directa o algoritmos basados en redes neurales complejas, los cuales podrían ser imposibles de inspeccionar. En segundo lugar, el resultado del algoritmo habría de ser predecible para que de esa manera la técnica psicoterapéutica sea una fuente de seguridad para el usuario, más que un motivo de incertidumbre. El algoritmo, por último, debería de ser robusto frente a la manipulación y debería de poderse determinar sin ambigüedades la persona humana responsable del mismo.

³⁹ Nicole M. Caulfield et al., «Exploring dissociation as a facilitator of suicide risk: A translational investigation using virtual reality», *J Affect Disord* 15, n.º 297 (2022): 517-24, <https://doi.org/doi:%2010.1016/j.jad.2021.10.097>.

⁴⁰ Peter Washington et al., «Data-Driven Diagnostics and the Potential of Mobile Artificial Intelligence for Digital Therapeutic Phenotyping in Computational Psychiatry», *Biol Psychiatry Cogn Neurosci Neuroimaging* 5, n.º 8 (2020): 759-69, <https://doi.org/10.1016/j.bpsc.2019.11.015>.

⁴¹ Matthew Bracher-Smith et al., «Machine learning for genetic prediction of psychiatric disorders: a systematic review», *Molecular Psychiatry* 26, n.º 1 (2021): 70-79, <https://doi.org/10.1038/s41380-020-0825-2>.

⁴² Amelia Fiske et al., «Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy», *Journal of Medical Internet Research* 21, n.º 5 (2019): 1-10, <https://doi.org/10.2196/13216>.

⁴³ Nick Bostrom y Elizier Yudkowsky, «The Ethics of Artificial Intelligence», en *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers (Oxford: Oxford University Press, 2021).

2. CIBERPRUDENCIA

Se han presentado en el primer apartado ocho posibles fuentes de cibervulnerabilidad, que invitan a una profunda reflexión sobre las implicaciones bioéticas que el pensamiento poshumanista plantea⁴⁴, donde la máquina y el humano se solapan en una suerte de nueva antropología androide⁴⁵ en espera de normatividad. Por otra parte, cada vez son más urgentes las demandas transhumanistas de protección frente a la probable emergencia de una superinteligencia artificial⁴⁶ que tendría una velocidad de computación y una base de conocimiento, en muchos órdenes de magnitud, superior a la capacidad humana. Llegado el caso, tal superinteligencia podría amenazar la supervivencia de un ser humano, como se ha mostrado, cibervulnerable en muchos aspectos.

En la exposición de este tercer apartado seguiremos el esquema que Gracia propone para el procedimiento deliberativo: 1. Deliberación sobre los hechos, 2. Deliberación sobre los valores, 3. Deliberación sobre los deberes, 4. Pruebas de consistencia de la decisión y 5. Toma de decisión definitiva⁴⁷. Se ubicarán los distintos tipos de cibervulnerabilidad en cada uno de estos momentos.

2.1. Deliberación sobre los hechos

Un hecho bien constatado es el efecto que le hiperconectividad digital tiene sobre la maduración del cerebro humano. La OMS ya ha advertido sobre ello, como se comentaba en el apartado de introducción. Las personas con trastornos afectivos, con adicciones o con patologías en el espectro autista parecen ser las más cibervulnerables. Por ello, es deseable una protección específica de sus neuroderechos, la cual supone también una apertura a nuevas oportunidades. Sería necesaria una mayor sensibilización y formación de los profesionales para atender a este grupo emergente de problemas clínicos relacionados con el mundo digital, que posiblemente esté siendo infradiagnosticado hoy en día. Sería también oportuna la recogida de variables conductuales y su uso para nutrir sistemas de IA —principalmente de ML— con intención diagnóstica. Es esta una oportunidad basada en tecnología ya consolidada que podría empezar a rendir resultados clínicos muy relevantes en la detección poblacional de sujetos de riesgo para trastornos neuropsiquiátricos, especialmente en el capítulo de los trastornos afectivos.

En los próximos cinco años es también previsible el desarrollo de sistemas de IA que mejoren la eficacia de las técnicas de neuroestimulación cerebral profunda y de estimulación vagal que actúan de manera inespecífica sobre el tono mental. Ello supone una valiosa oportunidad en el tratamiento de trastornos como la depresión o el Trastorno Obsesivo Compulsivo (TOC). Por otra parte, la inducción

⁴⁴ Luz María Pichardo, «Reinventing Bioethics in a Post-humanist and Post-truth Society. The Present and Future of Bioethics», *Persona y Bioética* 22, n.º 2 (2018): 212-22, <https://doi.org/10.5294/pebi.2018.22.2.2>.

⁴⁵ Martín Vargas Aragón, «Inteligencia artificial versus inteligencia natural: apuntes para una gnosología de fundamentación antropológica», *Kranion* 14, n.º 1 (2019): 24-30.

⁴⁶ Bostrom, *Superinteligencia. Caminos, peligros, estrategias*.

⁴⁷ Gracia, *El animal deliberante. Teoría y práctica de la deliberación moral*. (pp.645 y sig.).

de estados mentales alternativos está siendo explorada por la nueva disciplina de la psiquiatría psicodélica⁴⁸ y la IA se está utilizando en el diseño de nuevos medicamentos⁴⁹ en campos como la medicina del dolor, donde el uso de psicodélicos tiene un amplio desarrollo. Sería recomendable en los próximos años atender a las oportunidades y riesgos asociados a la inducción de estados alternativos de autoconsciencia con el apoyo de la IA. Considerando todo ello, nuestra primera recomendación prudencial es ante la hiperconectividad:

Recomendación 1: Seguir las recomendaciones de la OMS y de las sociedades científicas sobre el uso de tecnologías de la información y el conocimiento (TIC) en las distintas etapas de la vida y en las distintas condiciones de salud.

Y respecto a la alta cibervulnerabilidad específica de las personas con trastornos afectivos, adicciones o autismo podrían recomendarse las siguientes propuestas de prudencia ante la ciberpatología:

Recomendación 2: Implementar medidas específicas de ciberprevención en personas con patología psiquiátrica en el espectro de los trastornos afectivos, las adicciones o el autismo.

Recomendación 3: Auditar los algoritmos de los sistemas de ciberpsicoterapia para que sean transparentes, predecibles y robustos ante eventuales manipulaciones.

La confabulación de los sistemas de IA es un hecho fácil de comprobar. Basta con consultar en la última versión de ChatGPT, o en un sistema similar, sobre algún tema en el que el lector se considere experto. Es probable que la IA le devuelva resultados aparentemente veraces pero que, leídos en detalle, muestren falsedades que pudieran pasar desapercibidas a un lector inexperto en el tema, que es, precisamente, el tipo de lector preferente que acudirá a la IA para resolver sus dudas. La detección de conocimiento confabulado sólo es viable en un sistema inteligente que sea capaz de tener conocimiento sobre lo que sabe, pero también sobre lo que ignora. El usuario humano de IA puede valerse de su capacidad metacognitiva, la cual le permite diferenciar entre aquello en lo que puede considerarse experto, respecto a aquello otro en lo que es lego. La delimitación de la posible verdad anidada en una respuesta veraz ofrecida por la IA es competencia principal del usuario humano. En este cometido puede hacer uso de la docta ignorancia que Sócrates propusiera para el filósofo: «El Sócrates de Platón hizo célebre la tesis de la docta ignorancia, esto es, el enfoque según el cual la ignorancia puede contener algún aspecto propio de la sabiduría»⁵⁰. En ella se basa nuestra cuarta recomendación prudencial, ante la ilusión de conocimiento:

⁴⁸ Alex C. Kwan et al., «The neural basis of psychedelic action», *Nature Neuroscience* 25, n.º 11 (2022): 1407-19, <https://doi.org/10.1038/s41593-022-01177-4>.

⁴⁹ Jean Louis Kraus, «From “molecules of life” to new therapeutic approaches, an evolution marked by the advent of artificial intelligence: the cases of chronic pain and neuropathic disorders», *Drug Discovery Today* 26, n.º 4 (2021): 1070-75, <https://doi.org/10.1016/j.drudis.2021.01.009>.

⁵⁰ Marcelo D. Boeri, «La ignorancia socrática como virtud epistémica», *Pensamiento. Revista de Investigación e Información Filosófica* 77, n.º 293 (2021): 103-22, <https://doi.org/10.14422/pen.v77.i293.y2021.006>.

Recomendación 4: Antes de consultar a una IA sobre un tema que se desconoce, consultarle sobre un tema que el usuario conozca bien. Aplicar, a la respuesta que la IA ofrezca sobre el tema desconocido, el grado de credibilidad derivado de la respuesta dada al tema conocido.

2.2. *Deliberación sobre los valores*

El comportamiento de los animales no humanos está regulado por patrones instintivos heredables, que admiten cierto grado de aprendizaje ambiental. Por el contrario, la capacidad humana de valorar —su estimativa— depende en su principal medida del aprendizaje mediado por estímulos epigenéticos y ambientales. La neuroplasticidad mediada por las hormonas sexuales durante la adolescencia es clave en este proceso. Un sano neurodesarrollo pulsional es imprescindible para construir una adecuada estimativa personal, de lo cual deriva la siguiente recomendación prudencial ante el neurodesarrollo pulsional:

Recomendación 5: Educar a lo/as alumno/as adolescentes de enseñanzas medias para un adecuado uso de TIC en las que aparezcan contenidos sexuales, agresivos o relacionados con la alimentación y la imagen corporal, concienciándolos del impacto que la exposición a estos contenidos tendrá sobre su maduración estimativa.

La humanidad está entrando en una nueva era en la que la imagen de lo que significa ser humano cambiará con toda probabilidad. Cambios en la imagen de lo humano se han producido a lo largo de la historia y son motivo de estudio por la antropología filosófica, por ejemplo, el cambio que se produjo en la transición de la Edad Media al Renacimiento. Los problemas bioéticos derivados del posthumanismo, el transhumanismo y la mejora humana requerirán en las próximas décadas un abordaje riguroso y profundo por parte de la comunidad científica, jurídica, política y filosófica. En el corto plazo, habrá que atender a la manera en que los sistemas de IA modifican los valores en los que se basan la identidad de la población general. Los valores que sustentan la identidad natural y en los que se basa la ciberidentidad deberían de ser consistentes entre sí para evitar una disociación de la identidad perjudicial para la persona. La sexta es una recomendación de prudencia ante la ciberidentidad:

Recomendación 6: La comunidad natural y la comunidad virtual en las que la persona construye su identidad deberían de tener un alto grado de solapamiento para evitar una disociación de la identidad. Las personas deberían de poder rendir cuentas de su identidad virtual ante los miembros de su comunidad natural, y viceversa.

2.3. *Deliberación sobre los deberes*

Retomemos aquí lo expuesto más arriba sobre el primer tipo de ciberengaño, el propiciado por farsantes humanos. Entre las estrategias de los engañadores está la de hacer creer al estafado que tiene más poder de oportunidad o de conocimiento frente a un segundo agente. Dicho en breve, le inducen a «pasarse de listo»;

pensemos, por ejemplo, en el clásico «timo de la estampita». El ciberespacio es campo abonado para este tipo de engaño.

Parece, pues, prudente, que exista una estricta regulación estatal de actividades de alto riesgo accesibles a la población general, población que no tendría por qué tener conocimientos avanzados de IA. Este es el caso de la reciente regulación española del mercado de apuestas predictivas en plataformas como Polymarket o Kalshi en las que entran en juego de manera poco transparente la capacidad computacional de la intuición humana con la de los sistemas de IA⁵¹. La regulación estatal frente al ciberengaño compartiría justificación con la de la regulación frente a la modificación malintencionada del control apetitivo en los trastornos adictivos. En relación con ello proponemos tres recomendaciones sobre deberes de los Estados. Las dos primeras son de prudencia ante el ciberengaño:

Recomendación 7: Los Estados deberían promover en las enseñanzas medias la formación sobre oportunidades y riesgos de la IA.

Recomendación 8: Los Estados y agencias reguladoras deberían de controlar de manera estricta las actividades económicas, financieras o comerciales basadas en sistemas de IA, cuyos usuarios finales pudieran ser los ciudadanos medios.

Y la tercera es de prudencia ante la manipulación del control apetitivo:

Recomendación 9: Los Estados deberían perseguir las plataformas del metaverso que utilicen algoritmos susceptibles de inducir el desarrollo de adicciones comportamentales en los usuarios.

2.4. Pruebas de consistencia de la decisión

El premio Nobel de Economía de 1978, Herbert Simon, entiende irónicamente por racionalidad olímpica aquella que sería exclusiva de los dioses, que contemplan directamente la verdad, y a la que aspirarían desatinadamente aquellos que siguen ciegamente la teoría de la elección racional⁵². Puede entenderse que la singularidad ideada por Bostrom⁵³ no sería sino una reificación teológica de la perfecta racionalidad olímpica irónicamente imaginada décadas antes por Simon, quien abrió la senda de una teoría de la racionalidad que desde los años 50 del siglo XX ha acreditado numerosas virtudes y que podría ser también de utilidad como marco de ciberprudencia. Se trata de la teoría de la racionalidad acotada —*Bounded Rationality* en inglés—⁵⁴, caracterizada, entre otros atributos, por tener en cuenta la capacidad de cómputo del agente epistémico, el tiempo de computación de que dispone para realizar la computación y la cantidad de conocimiento a la que puede

⁵¹ Cf. Ministerio de Derechos Sociales, Consumo y Agenda 2030, «Consumo abre expediente sancionador a las plataformas Polymarket y Kalshi y ordena el bloqueo de sus webs», nota de prensa, 26 de mayo de 2026, acceso el 31 de mayo de 2026, <https://www.dsca.gob.es/es/comunicacion/notas-prensa/consumo-abre-expediente-sancionador-plataformas-polymarket-kalshi-ordena>.

⁵² Herbert A. Simon, «A Behavioral Model of Rational Choice», *The Quarterly Journal of Economics* 69, n.º 1 (1955): 99-118; Herbert A. Simon, *Naturaleza y límites de la razón humana*. (Fondo de Cultura Económica, 1989).

⁵³ Bostrom, *Superinteligencia. Caminos, peligros, estrategias*.

⁵⁴ Gregory Wheeler, «Bounded Rationality», *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), ed. Edward N. Zalta 2020, <https://doi.org/10.1145/544317.544327>.

acceder en ese tiempo. Frente a la superinteligencia teologizada en la singularidad, cabe recordar con Epicuro el desinterés de los dioses hacia las cuestiones humanas para reivindicar las virtudes de la racionalidad acotada⁵⁵. Bien pudiera ser esta una herramienta de prudencia, en nuestro caso ciberprudencia, ante a la olímpica racionalidad de la eventual singularidad. Una herramienta cercana a la prudente sabiduría, que junto a la capacidad computacional atienda a la inteligencia emocional, la moralidad y la empatía⁵⁶, tratando con justicia epistémica a todos los agentes en interacción. Las decisiones tomadas al amparo de este marco tendrían una consistencia máxima al implicar todos los aspectos cognitivos de los agentes decisores, y no solo una parte de ellos.

José Francisco Álvarez aporta argumentos alineados con la importancia de atender a las valoraciones, y no sólo a los hechos, en los procesos de toma de decisiones. Advierte sobre la falacia subyacente a la idea de que la toma racional de decisiones, incluso en ciencia, esté libre de valoraciones⁵⁷. Por el contrario, al poder incorporar presupuestos éticos y, en general, valorativos, la racionalidad acotada tendría la virtud de aunar en una sola teoría la toma de decisiones, que resultarían por ello decisiones racionales a la vez que prudentes. Así la teoría de la racionalidad acotada resulta especialmente apropiada para una eventual modelización de la ciberprudencia.

En la tradición de racionalidad acotada iniciada por Simon, otro laureado con el Nobel de Economía, esta vez en 2002, el psicólogo David Kahneman, describe en los seres humanos la existencia de dos sistemas para la toma de decisiones⁵⁸. El Sistema 1 es rápido y toma decisiones automatizadas en situaciones de alta carga emocional, como puedan ser las situaciones de peligro. Por su parte, el Sistema 2 toma con esfuerzo decisiones lentas y sopesadas racionalmente que, a su vez, controlan las decisiones tomadas por el Sistema 1. Sopesar racionalmente la toma de decisiones es, precisamente, el objetivo de la deliberación. Pero nótese que, frente al mero uso de la razón que puede hacerse aisladamente al modo del viejo Kant, la deliberación es una práctica discursiva y por lo tanto grupal. Karl Otto Apel⁵⁹ ha señalado la importancia del uso público de la razón, de manera que el significado de las proposiciones ha de ser siempre público, ha de ser una razón discursiva.

2.5. Toma de decisión definitiva

Adela Cortina recomienda una forma concreta de razón deliberativa aunada con las virtudes democráticas que bautiza como democracia comunicativa. Apelando a Habermas, escribe Cortina sobre las normas:

⁵⁵ Gerd Gigerenzer, «Fast and Frugal Heuristics: The Tools of Bounded Rationality», en *Blackwell handbook of judgment and decision making*, en D. Koehler y N. Harvey (Malden: Blackwell Publishing, 2004).

⁵⁶ Lee et al., «Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom».

⁵⁷ J. Francisco Álvarez, «Racionalidad, modelos humanos y economía normativa», *Argumentos de razón técnica* 3 (2000): 93-114.

⁵⁸ Daniel Kahneman, *Pensar rápido, pensar despacio* (Madrid: Debate, 2012).

⁵⁹ Karl Otto Apel, *La transformación de la filosofía* (Madrid: Taurus, 1985).

...la única forma de que merezcan respeto es que los destinatarios sean a la vez sus autores; lo cual significa que cualquier norma jurídica tiene que poder ser justificada por medio del discurso, tiene que poder ser aceptada racionalmente por sus destinatarios⁶⁰.

Parece esta una buena guía para la construcción de una racionalidad supervisora, un Sistema 2 de Kahneman asentado, no ya en el cerebro de un solo ser humano, sino en toda una comunidad dialogante, incluso a escala mundial, e incluso incorporando el diálogo con las IAs. Se trataría de utilizar una forma de racionalidad acotada que considere tanto los aspectos valorativos como los hechos para construir comunitariamente normas válidas que sean útiles en la toma de decisiones a la vez racionales y prudentes.

La razón cordial en el sentido orteguiano puede constituirse en la necesaria racionalidad acotada supervisora del metaverso. Una razón cordial que defienda al ser humano frente a las amenazas derivadas de la cibervulnerabilidad de nuestra especie, una razón cordial que sea capaz de construir una urdimbre metafórica⁶¹ de la que participen agentes epistémicos tanto naturales como artificiales. Tal es la forja del carácter deseable para los seres humanos del siglo XXI, una ciudadanía educada para aprovechar las virtudes del corazón, de la emoción y de la razón cordial⁶² en alianza con los fríos cálculos de la inteligencia artificial. Por ello, frente al riesgo de injusticia epistémica hacia los humanos, cabría proponer el derecho a una jerarquía epistémica de los juicios y valoraciones humanas sobre la toma de decisiones ejecutadas por IA, pues las decisiones que equilibran los elementos emocionales y los racionales son idóneas cuando permiten participar de ellas a todos sus autores. Concluye, con esto, nuestra última recomendación, que es de prudencia ante la injusticia epistémica:

Recomendación 10: Priorizar las decisiones deliberadas en las comunidades humanas naturales, frente a la toma de decisiones basadas en la decisión de agentes epistémicos de IA.

REFERENCIAS

- Álvarez, J. Francisco. «Racionalidad, modelos humanos y economía normativa». *Argumentos de razón técnica* 3 (2000): 93-114.
- Anderson, Craig A., Brad J. Bushman, Bruce D. Bartholow, et al. «Screen violence and youth behavior». *Pediatrics* 140, n.º November (2017): S142-47. <https://doi.org/10.1542/peds.2016-1758T>.
- Apel, Karl Otto. *La transformación de la filosofía*. Madrid: Taurus, 1985.

⁶⁰ Adela Cortina, «Ciudadanía democrática: ética, política y religión. XIX Conferencias Aranguren», *Isegoría* 0, n.º 44 (2011): 13-55, <https://doi.org/10.3989/isegoria.2011.i44.718>. (p.31)

⁶¹ Tomás Domingo Moratalla, «Ortega y Gasset: cuando la metáfora da (mucho) que pensar. La urdimbre metafórica de la razón cordial», *Red de investigaciones filosóficas José Sanmartín Esplugues*, 2022, <https://proyectoscio.ucv.es/articulos-filosoficos/ortega-y-gasset-metaphora/>.

⁶² Adela Cortina Orts, *Ética de la razón cordial: educar en la ciudadanía en el siglo XXI* (Oviedo: Ediciones Nobel, 2007); Javier Sánchez Pachón, «Adela Cortina: el reto de la ética cordial», *Brocar. Cuadernos de Investigación Histórica*, n.º 39 (noviembre de 2015): 397-422, <https://doi.org/10.18172/brocar.2901>.

- Boeri, Marcelo D. «La ignorancia socrática como virtud epistémica». *Pensamiento. Revista de Investigación e Información Filosófica* 77, n.o 293 (2021): 103-22. <https://doi.org/10.14422/pen.v77.i293.y2021.006>.
- Bostrom, Nick. *Superinteligencia. Caminos, peligros, estrategias*. Zaragoza: Teell Editorial, 2014.
- Bostrom, Nick y Eliezer Yudkowsky. «The Ethics of Artificial Intelligence». En *Philosophy of Mind: Classical and Contemporary Readings*, editado por David J. Chalmers. Oxford: Oxford University Press, 2021.
- Bracher-Smith, Matthew, Karen Crawford, y Valentina Escott-Price. «Machine learning for genetic prediction of psychiatric disorders: a systematic review». *Molecular Psychiatry* 26, n.o 1 (2021): 70-79. <https://doi.org/10.1038/s41380-020-0825-2>.
- Bringsjord, Selmer y Naveen Sundar Govindarajulu. «Artificial Intelligence». *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition). Editado por Edward N. Zalta y Uri Nodelman.
- Caulfield, Nicole M., Aleksandr T. Karnick, y Daniel W. Capron. «Exploring dissociation as a facilitator of suicide risk: A translational investigation using virtual reality». *J Affect Disord* 15, n.o 297 (2022): 517-24. <https://doi.org/doi:%2010.1016/j.jad.2021.10.097>.
- Clark, Luke, y Martin Zack. «Engineered highs: Reward variability and frequency as potential prerequisites of behavioural addiction». *Addictive Behaviors* 140, n.o December 2022 (2023): 107626. <https://doi.org/10.1016/j.addbeh.2023.107626>.
- Cortina, Adela. «Ciudadanía democrática: ética, política y religión. XIX Conferencias Aranguren». *Isegoría* 0, n.o 44 (2011): 13-55. <https://doi.org/10.3989/isegoria.2011.i44.718>.
- Cortina Orts, Adela. *Ética de la razón cordial: educar en la ciudadanía en el siglo XXI*. Oviedo: Ediciones Nobel, 2007.
- De Jaegher, Hanne, Anssi Peräkylä, y Melisa Stevanovic. «The co-creation of meaningful action: bridging enaction and interactional sociology». *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, n.o 1693 (2016): 20150378. <https://doi.org/10.1098/rstb.2015.0378>.
- Domingo Moratalla, Tomás. «Ortega y Gasset: cuando la metáfora da (mucho) que pensar. La urdimbre metafórica de la razón cordial». *Red de investigaciones filosóficas José Sanmartín Esplugues*, 2022. <https://proyectoscio.ucv.es/articulos-filosoficos/ortega-y-gasset-metфора/>.
- Dresp-Langley, Birgitta y Axel Hutt. «Digital Addiction and Sleep». *International Journal of Environmental Research and Public Health* 19, n.o 11 (2022): 2-19. <https://doi.org/10.3390/ijerph19116910>.
- Fiske, Amelia, Peter Henningsen y Alena Buyx. «Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy». *Journal of Medical Internet Research* 21, n.o 5 (2019): 1-10. <https://doi.org/10.2196/13216>.
- Gigerenzer, Gerd. «Fast and Frugal Heuristics: The Tools of Bounded Rationality». En *Blackwell Handbook of Judgment and Decision Making*, editado por Derek J. Koehler y Nigel Harvey. Malden: Blackwell Publishing, 2004.
- Gracia, Diego. *El animal deliberante. Teoría y práctica de la deliberación moral*. Madrid: Triacastela, 2025.
- Graham, Sarah, Colin Depp, Ellen E. Lee, et al. «Artificial Intelligence for Mental Health and Mental Illnesses: an Overview». *Current Psychiatry Reports* 21, n.o 11 (2019): 116. <https://doi.org/10.1007/s11920-019-1094-0>.
- Hebb, Donald O. *The organization of behavior*. Nueva York: Wiley, 1949.
- Ioannidis, Konstantinos y Samuel R. Chamberlain. «Digital Hazards for Feeding and Eating: What We Know and What We Don't». *Current Psychiatry Reports* 23, n.o 9 (2021): 2-8. <https://doi.org/10.1007/s11920-021-01271-7>.

- Jan, Zainab, Noor AI-Ansari, Osama Mousa, et al. «The role of machine learning in diagnosing bipolar disorder: Scoping review». *Journal of Medical Internet Research* 23, n.o 11 (2021): 1-17. <https://doi.org/10.2196/29749>.
- Jaureguizar, Joana, Iratxe Redondo, Juan Manuel Machimbarrena y Sebastian Wachs. «Editorial: Risks of “cyber-relationships” in adolescents and young people». *Frontiers in Psychology* 14, n.o 1 (2023): 4-6. <https://doi.org/10.3389/fpsyg.2023.1118736>.
- Ji, Ziwei, Nayeon Lee, Rita Frieske, et al. «Survey of Hallucination in Natural Language Generation». *ACM Computing Surveys* 55, n.o 12 (2023): 1-38. <https://doi.org/10.1145/3571730>.
- Johannes, Schiebener, Laier Christian, y Brand Matthias. «Getting stuck with pornography? Overuse or neglect of cybersex cues in a multitasking situation is related to symptoms of cybersex addiction». *Journal of Behavioral Addictions* 4, n.o 1 (2015): 14-21. <https://doi.org/10.1556/JBA.4.2015.1.5>.
- Kahneman, Daniel. *Pensar rápido, pensar despacio*. Madrid: Debate, 2012.
- Kato, Takahiro A., Shigenobu Kanba y Alan R. Teo. «Hikikomori: Multidimensional understanding, assessment, and future international perspectives». *Psychiatry and Clinical Neurosciences* 73, n.o 8 (2019): 427-40. <https://doi.org/10.1111/pcn.12895>.
- Kim, Jooyoung. «Advertising in the Metaverse: Research Agenda». *Journal of Interactive Advertising* 21, n.o 3 (2021): 141-44. <https://doi.org/10.1080/15252019.2021.2001273>.
- Korte, Martin. «The impact of the digital revolution on human brain and behavior: Where do we stand?». *Dialogues in Clinical Neuroscience* 22, n.o 2 (2020): 101-11. <https://doi.org/10.31887/DCNS.2020.22.2/mkorte>.
- Kraus, Jean Louis. «From “molecules of life” to new therapeutic approaches, an evolution marked by the advent of artificial intelligence: the cases of chronic pain and neuropathic disorders». *Drug Discovery Today* 26, n.o 4 (2021): 1070-75. <https://doi.org/10.1016/j.drudis.2021.01.009>.
- Kwan, Alex C., David E. Olson, Katrin H. Preller y Bryan L. Roth. «The neural basis of psychedelic action». *Nature Neuroscience* 25, n.o 11 (2022): 1407-19. <https://doi.org/10.1038/s41593-022-01177-4>.
- Lee, Ellen E., John Torous, Munmun De Choudhury, et al. «Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom». *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 6, n.o 9 (2021): 856-64. <https://doi.org/10.1016/j.bpsc.2021.02.001>.
- León XIV. Magnífica humanidad. Carta encíclica sobre la custodia de la persona humana en el tiempo de la inteligencia artificial. 15 de mayo de 2026. <https://www.vatican.va/content/leo-xiv/es/encyclicals/documents/20260515-magnifica-humanitas.html>.
- Liu, Yang, Janne Kauttonen, Bowen Zhao, Xiaobai Li, y Wei Peng. «Editorial: Towards Emotion AI to next Generation Healthcare and Education». *Frontiers in Psychology* 15 (diciembre de 2024): 1533053. <https://doi.org/10.3389/fpsyg.2024.1533053>.
- McCadden, Melissa, Katrina Hui, y Daniel Z. Buchman. «Evidence, ethics and the promise of artificial intelligence in psychiatry». *Journal of Medical Ethics* 0 (2022): 1-7. <https://doi.org/10.1136/jme-2022-108447>.
- Morán Espinosa, Alejandra. «Responsabilidad penal de la Inteligencia Artificial (IA). ¿La próxima frontera?». *IUS. Revista del Instituto de Ciencias Jurídicas de Puebla, México* 15, n.o 48 (2021): 289-323. <https://theconversation.com/mi-ciruja->.
- O'Dell, Bessie, Katherine Stevens, Anneka Tomlinson, Ilna Singh y Andrea Cipriani. «Building trust in artificial intelligence and new technologies in mental health». *Evidence-Based Mental Health* 25, n.o 2 (2022): 45-46. <https://doi.org/10.1136/ebmental-2022-300489>.
- Passos, Ives Cavalcante, Pedro Ballester, Francisco Diego Rabelo-da-Ponte y Flavio Kapczinski. «Precision Psychiatry: The Future Is Now». *Canadian Journal of Psychiatry* 67, n.o 1 (2022): 21-25. <https://doi.org/10.1177/0706743721998044>.

- Pfeifer, Jennifer H. y Sarah Jayne Blakemore. «Adolescent social cognitive and affective neuroscience: Past, present, and future». *Social Cognitive and Affective Neuroscience* 7, n.o 1 (2012): 1-10. <https://doi.org/10.1093/scan/nsr099>.
- Pichardo, Luz María. «Reinventing Bioethics in a Post-humanist and Post-truth Society. The Present and Future of Bioethics». *Persona y Bioética* 22, n.o 2 (2018): 212-22. <https://doi.org/10.5294/pebi.2018.22.2.2>.
- San Martín, Javier. Creencia y realidad: Desde un diálogo con Ortega y sus intérpretes. Madrid: Editorial Fundación Sicómoro, 2025.
- Sánchez Pachón, Javier. «Adela Cortina: el reto de la ética cordial». *Brocar. Cuadernos de Investigación Histórica*, n.o 39 (noviembre de 2015): 397-422. <https://doi.org/10.18172/brocar.2901>.
- Simon, Herbert A. «A Behavioral Model of Rational Choice». *The Quarterly Journal of Economics* 69, n.o 1 (1955): 99-118.
- Simon, Herbert A. *Naturaleza y límites de la razón humana*. México: Fondo de Cultura Económica, 1989.
- Swing, Edward L., Douglas A. Gentile, Craig A. Anderson y David A. Walsh. «Television and video game exposure and the development of attention problems». *Pediatrics* 126, n.o 2 (2010): 214-21. <https://doi.org/10.1542/peds.2009-1508>.
- Taddeo, Mariarosaria y Luciano Floridi. «The case for e-trust». *Ethics and Information Technology* 13, n.o 1 (2011): 1-3. <https://doi.org/10.1007/s10676-010-9263-1>.
- Tang, J., A. LeBel, S. Jain y AG Huth. «Semantic reconstruction of continuous language from non-invasive brain recordings». *Nat Neurosci* 26 (2023): 858-66. <https://doi.org/https://doi.org/10.1038/s41593-023-01304-9>.
- Usmani, Sadia Suhail, Medha Sharath y Meghana Mehendale. «Future of mental health in the metaverse». *General Psychiatry* 35, n.o 4 (2022): 1-6. <https://doi.org/10.1136/gp-sych-2022-100825>.
- Vargas Aragón, Martín. «Inteligencia artificial versus inteligencia natural: apuntes para una gnoseología de fundamentación antropológica». *Kranion* 14, n.o 1 (2019): 24-30.
- Walsh, Jessica J., Daniel J. Christoffel, Xiaoting Wu, Matthew B. Pomrenze y Robert C. Malenka. «Dissecting neural mechanisms of prosocial behaviors». *Current Opinion in Neurobiology* 68 (2021): 9-14. <https://doi.org/10.1016/j.conb.2020.11.006>.
- Washington, Peter, Natalie Park, Parishkrita Srivastava, et al. «Data-Driven Diagnostics and the Potential of Mobile Artificial Intelligence for Digital Therapeutic Phenotyping in Computational Psychiatry». *Biol Psychiatry Cogn Neurosci Neuroimaging* 5, n.o 8 (2020): 759-69. <https://doi.org/10.1016/j.bpsc.2019.11.015>.
- Wheeler, Gregory. «Bounded Rationality». The Stanford Encyclopedia of Philosophy (Fall 2020 Edition). Editado por Edward N. Zalta. 2020. <https://doi.org/10.1145/544317.544327>.

Facultad de Medicina de la Universidad de Valladolid
e Instituto Torozos de Iatrofilosofía.
Email: martinlorenzo.vargas@uva.es
ORCID: 0000-0002-3175-0643

MARTÍN L. VARGAS-ARAGÓN

[Artículo aprobado para publicación en diciembre de 2025]