**Daimon. Revista Internacional de Filosofía**, nº 90 (2023), pp. 99-114
ISSN: 1130-0507 (papel) y 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.562821


# Ética digital discursiva: de la explicabilidad a la participación*

## Discursive digital ethics: From explicability to participation

*DOMINGO GARCÍA-MARZÁ***

**Resumen.** el presente artículo tiene como objetivo presentar los rasgos básicos de una ética digital dialógica a partir de una lectura crítica del documento elaborado por el grupo independiente de expertos de alto nivel para la Comisión Europea *Ethics Guidelines for Trustworthy AI* (High-level expert Group on Artificial Intelligence, 2019). Una ética digital que tiene en el diálogo y acuerdo posible de todos los agentes implicados y afectados por la realidad digital su horizonte normativo de actuación, su criterio de justicia. La finalidad es mostrar que, en su esfuerzo por generar una voluntad común y una gobernanza europeas ante la actual revolución industrial, la participación de todas las partes implicadas no solo es recomendable sino moralmente exigible. El reconocimiento de la igual dignidad que implica una Inteligencia Artificial centrada en las personas no es ni siquiera pensable sin el horizonte de una participación igual. Sin ella, la confianza no puede generarse ni garantizarse. Como pretendemos mostrar, en este objetivo juega un papel decisivo el nuevo principio de explicabilidad, principio que posee un valor moral y no solo instrumental.

**Abstract.** This article is intended to present a proposal for dialogic digital ethics on a critical reading of the European Commission's independent high-level expert group's document *Ethics Guidelines for Trustworthy AI* (2019). These would be digital ethics with a normative horizon for action and criteria for justice based on dialogue and possible agreement between all agents involved and affected by the digital reality. The aim is to show that the participation of all parties involved is not merely advisable but morally required as part of the Commission's effort to generate common European willingness and governance to deal with the fourth industrial revolution now going on. The acknowledgement of equal dignity implied by people-centric artificial intelligence (AI) is utterly unthinkable without this possibility of equal participation. Without it, trust cannot be generated or guaranteed. As we intend to show, the new principle of explicability plays a decisive role in this objective as a principle with a moral as well as an instrumental.

---

Recibido: 28/03/2023. Aceptado: 04/06/2023.

* Este trabajo se enmarca dentro de los objetivos del Proyecto de Investigación Científica y Desarrollo Tecnológico «Ética aplicada y confiabilidad para una Inteligencia Artificial» [PID2019- 109078RB-C21] financiado por el Ministerio de Ciencia e Innovación, así como en las actividades del grupo de investigación de excelencia CIPROM/2021/072 de la Comunitat Valenciana.

** Catedrático de Ética y Filosofía Política en la Universitat Jaume I de Castellón. Es autor, entre otros, de los siguientes libros: *Public reason and Applieds Ethics* (junto con A. Cortina y J. Conill (eds.) Londres, 2008); *Ética y Filosofia Política. Homenaje a Adela Cortina* (junto con J. Félix Lozano; E. Martinez y J. C. Siurana, Madrid, 2018). Autor de numerosos artículos sobre la relación entre ética, política y economía y su aplicación en el diseño institucional. Los resultados de estas investigaciones se han plasmado en diversas instituciones públicas y privadas. Es co-director del Programa interuniversitario de Doctorado «Ética y Democracia» en la Universitat Jaume I y patrono de la Fundación ETNOR.

Con este fin este trabajo se estructura en tres partes. En primer lugar, se argumentará la propuesta de una ética digital dialógica encargada, como ética aplicada, de explicitar las bases éticas que subyacen a la confianza en la Inteligencia artificial, en sus decisiones, prácticas e instituciones. Desde este marco ético, en segundo lugar, se analizarán las Directrices Europeas y se propondrá revisar la consideración de la inclusión y participación de todas las partes implicadas no solo como un requisito recomendable sino como una exigencia moral, destacando la necesidad de justificar y potenciar una participación real y efectiva. Una justificación que se realizará, ya en el tercer punto, desde el principio de explicabilidad como principio moral, siguiente el camino del principio kantiano de publicidad. La finalidad es avanzar, desde esta propuesta de una ética digital discursiva, un diseño institucional capaz de responder de esta exigencia moral de la participación libre e igual de todos los afectados e implicados. El *principio de explicabilidad* se convierte así en un principio básico para garantizar este saber moral para la toma de decisiones y la creación de espacios de confianza "dentro" de las instituciones que conforman el sistema socio-técnico de la Inteligencia artificial.
**Palabras clave:** Ética, Ética aplicada, Ética digital, Inteligencia Artificial, recursos morales, participación, diseño institucional, infraestructura ética.

To this purpose, this study is structured in three parts. First, it will argue the proposal of a dialogic digital ethics in charge, as applied ethics, of making explicit the ethical bases that underlie the trust in Artificial Intelligence, in its decisions, practices and institutions. From this ethical framework, secondly, the European Guidelines will be analyzed, highlighting the need to justify and enhance the participation of all parties involved. This justification will be based on the Kantian principle of publicity. Finally, from this proposal of a discursive digital ethics, an ethical infrastructure capable of integrating all institutional design and all algorithmic development with this moral requirement of free and equal participation of all those affected and involved will be proposed. The principle of explicability thus becomes a basic principle to guarantee this moral knowledge for decision-making and the creation of spaces of trust "within" the institutions that make up the socio-technical system of Artificial Intelligence.
**Keywords:** Ethics, applied ethics, Digital ethics, artificial intelligence, moral resources, institutional design, participation, ethical infrastructure.

## 1. Hacia una ética digital dialógica como ética aplicada

Estas directrices éticas, aunque llegan con retraso dada la aceleración tecnológica actual, siempre son bienvenidas en su función de orientar la toma de decisiones individuales e institucionales y para influir en el desarrollo legislativo. Mientras que nuestros gobernantes insisten y repiten por doquier que todo avance tecnológico debe ir acompañado de una visión ética, de una base humanista, de un proyecto de progreso, de un desarrollo innovador e inclusivo, etc., las estrategias, las políticas públicas, las grandes tecnológicas, los contratos público-privados, etc., la realidad, en suma, lo desmiente. Un caso práctico nos puede servir de ejemplo. Lo encontramos en la *Estrategia de Inteligencia Artificial de la Comunidad Valenciana*, donde podemos leer: "los beneficios de aplicar la IA se reflejarán en una mejor toma de decisiones al tener en cuenta todas las posibles variables y contrastar un mayor número de datos. Al estar basadas en algoritmos que no se ven afectados por subjetividades personales, las decisiones son más objetivas. También son más rápidas porque la capacidad de cómputo supera a la capacidad de la inteligencia humana (Generalitat Valenciana, 2018: 5).

Si no se denuncian estos sesgos de superioridad, neutralidad y objetivismo, que parecen acompañar a la comprensión de la actual transformación digital, si no se desvelan estos prejuicios, una ética digital —el análisis de lo correcto o incorrecto en el ámbito tecnológico y digital— solo podrá aplicarse después de aparecido el problema, tras las consecuencias de las decisiones y acciones ya tomadas desde esta racionalidad tecnológica reducida a los datos y macrodatos, a los algoritmos y a la inteligencia artificial. Y ya será, como la experiencia nos muestra en el día a día, demasiado tarde (Apel, 1988; Rehg, 2015; Yuste et al., 2017).

Esta negativa a reconocer cualquier intervención humana, como si los datos estuvieran "ahí fuera" esperando ser descubiertos, como si los algoritmos no fueran de facto fruto de nuestra interpretación de la realidad, de nuestros intereses de conocimiento, no solo demuestra una clara torpeza para comprender una situación, sino también una clara intencionalidad para *cosificar*, y, por tanto, *obstruir* toda posibilidad de diseño y gobernanza éticos (García-Marzá, 2022). Esta denuncia es la primera tarea de una ética digital dialógica, pero no la única como a continuación veremos.

Una ética digital, centrada en las bases morales que subyacen a la confianza que depositamos en las diferentes prácticas digitales y sus respectivas tecnologías, debe comenzar por criticar estos prejuicios que, una vez más, reaparecen vinculados al dominio tecnológico que la ciencia presupone, ya sea como cientificismo, objetivismo, positivismo, etc. y que hoy son recuperados por las neurociencias y la Inteligencia Artificial (García-Marzá, 2019). El punto de partida para una ética digital no puede ser otro que la superación de este obstáculo, de este dogmatismo epistemológico que impide abordar la parte humana que define y gestiona todo tipo de tecnologías. Si la decisión algorítmica es más justa que la decisión humana, tanto la autonomía moral como la política desaparecen. Debemos mostrar de nuevo, insistir sin cansarnos, que la definición, la identificación y selección, de un dato depende siempre de determinados intereses y, por lo tanto, cuando se utilizan para la construcción de los algoritmos, está ya cargados de valores (Mittelstad et. al.2016; Floridi, 2019).

De ahí que la Comisión Europea, en su informe *Generar confianza en la IA centrada en el ser humano* (COM.2019), afirme que "La IA trae retos, ya que permite a las máquinas "aprender" y tomar decisiones y ejecutarlas sin intervención humana. Ahora bien, las decisiones adoptadas por algoritmos pueden dar datos incompletos y, por tanto, no fiables, que puede ser manipulados, sesgados o simplemente estar equivocados" (COM.2019: 2). Como el resto de tecnologías, las digitales son un instrumento, un medio, para conseguir determinados fines, no un fin en sí mismo. Como veremos, el significado de un dato, por qué un dato lo es y otro no, siempre es una decisión humana. Los algoritmos dependen de una realidad construida desde un interés determinado. Desde este interés se construye el algoritmo que reunirá, integrará y dará sentido a los datos. Las redes y el internet de las cosas son la fuente de los macrodatos, los algoritmos ordenan estos datos (García-Marzá; Calvo, 2022).

El dictamen 4/2015 del European Data Protection Supervisor titulado *Hacia una nueva Ética Digital. Datos, Dignidad y Tecnología* (2015), apuesta por estimular un debate abierto y documentado sobre la definición, justificación y aplicación de una *nueva ética digital* (2015: 5). Un debate en el que participen la sociedad civil, los diseñadores, las empresas, los académicos, las autoridades reguladoras, etc. Una ética que permita "mejorar los beneficios de la tecnología para la sociedad y la economía por vías que refuercen los derechos y las libertades de las personas físicas" (European Data Protection Supervisor, 2015: 5).

Perfecto, podríamos pensar, pero si no se concreta esta participación, si no se establecen los mecanismos y procedimientos institucionales que la posibiliten, solo son vanas palabras. La duda es siempre la misma: ¿no estamos ante un nuevo intento para encubrir las injusticias provocadas por las nuevas tecnologías digitales y las grandes empresas tecnológicas, dueñas hoy por hoy de la globalización, con la piel de cordero de la dignidad humana? De hecho, este peligro de caer en un *ethics washing* está claramente explícito en el reciente informe de la UNESCO (2022).

Como ética aplicada, una ética digital dialógica tiene como objetivo explicitar y gestionar las bases éticas de la confianza depositada en la llamada revolución digital. Se concibe como una ética aplicada porque su objetivo no se detiene en la explicitación del saber moral que utilizamos en todo proceso de digitalización, desde la determinación de los datos y su integración a través de los algoritmos, hasta el aprendizaje autónomo y la Inteligencia Artificial, pasando por la hiperconectividad proporcionada por el internet de las cosas, las máquinas y robots, así como los problemas derivados de la computación en nube. El bien primario que aporta la práctica digital son precisamente los datos como significados atribuidos a los signos, como interpretaciones de lo dado, de la realidad. Los datos no recogen la realidad, lo dado, sino aquello que nos interesa de la misma y no lo hacen desde la lógica de la causa efecto, sino desde la lógica de la correlación, de las tendencias y patrones (García-Marzá et al., 2004). La neutralidad queda fuera de esta lógica, pues siempre se trata de una "elección" definir un hecho dado como un dato, para después afirmar que este es, por ejemplo, el comportamiento humano.

Podemos hablar también de ética de la inteligencia artificial, de ética algorítmica, de ética de datos, etc., pero en todos estos casos la preocupación y el interés es el mismo: dar razón de la significación y el valor que tiene hoy en día, en contextos globales sin orden jurídico global, la dimensión ética. Una dimensión que se caracteriza, no lo olvidemos, por su pretensión de universalidad. De hecho, hoy en día, la ética digital, con sus valores, principios, directrices, etc., ya interviene en el mundo de la tecnología mucho más que cualquier otra fuerza, pues la percepción y valoración de lo que es moralmente bueno, correcto o justo, influye en la opinión pública, en lo socialmente aceptable o preferible y en lo políticamente factible, y por tanto, en última instancia, en lo legalmente exigible. Por desgracia, este poder de intervención solo suele aparecer con las consecuencias negativas producidas o esperadas, esto es, cuando el mal está hecho.

La ética digital se entiende como una ética aplicada cuyo ámbito de acción es la práctica digital, sus procesos y tecnologías; sus sistemas de toma de decisiones, así como los marcos institucionales en los que se producen —empresas, centros de investigación, etc. Floridi se refiere a la *infoesfera* como conjunto de prácticas conducidas por y dependientes de los datos, ocupándose la ética digital del alcance y las reglas que permiten las interacciones en esta nueva esfera digital. El problema ya no es tanto la innovación digital, como su gobernanza (Floridi,2018; Calvo, 2021).

## 2. La generación de confianza: una revisión del marco europeo para una IA confiable

Al igual que ocurre con el resto de las éticas aplicadas, la generación y el mantenimiento de la confianza en la esfera digital requiere tres pasos básicos derivados de la imposibilidad

de trasladar automáticamente los principios éticos a la práctica, pues no dejan de ser obligaciones morales abstractas (García-Marzá, 2004). De ahí que necesitemos diferenciar tres niveles en el camino que va de la obligación moral a su realización práctica:

- *Nivel de justificación:* si queremos hablar de validez moral y, por tanto, de normatividad, de lo justo o correcto, debemos fundamentar los principios morales con los que explicitamos nuestro saber moral. La fundamentación debe apelar a razones que justifiquen la universalidad de los principios, esto es, que garanticen la igual dignidad que nos define como personas. Desde este punto de vista, los principios éticos son condiciones de posibilidad de esta igual dignidad.
- *Nivel de realización:* este saber moral debe transformarse en recursos, capacidades y competencias que, dentro de las instituciones, *todo* ser humano tiene a su disposición a la hora de relacionarse con los demás. Estos recursos morales solo aparecen cuando consideramos a los demás como iguales, como interlocutores válidos, cuando actuamos desde el reconocimiento recíproco siguiendo un interés que es de todos, no solo de unos cuantos o de uno solo (García-Marzá, 2004).
- *Nivel de concreción organizativa:* se requiere tanto una cultura como una infraestructura ética para que estos recursos morales puedan ser utilizados en todas las fases que supone la construcción del espacio digital y no solo, como veremos, cuando el problema ya ha aparecido. En este sentido es importante centrarnos en el diseño institucional, tanto de los procesos digitales como de las organizaciones que se encargan de la investigación y de su producción. La confianza en las organizaciones que realizan investigación e innovación, así como las encargadas de su financiación, depende de las razones que posean para sostener su credibilidad y su reputación. Unas razones que exigen la presencia de espacios de participación, de espacios capaces de generar confianza, en el interior de las mismas, en su estructura organizativa. De ahí que hablemos de un diseño institucional para la aplicación y el desarrollo de una ética digital dialógica, de la necesidad de una infraestructura ética (García-Marzá, 2017).

La misma idea de que no es posible pasar directamente de los principios éticos a las prácticas digitales necesitadas de regulación la encontramos en la propuesta *Ethics Guidelines for Trustworthy AI* (High-level expert Group on Artificial Intelligence, 2019) promovida por la Comisión Europea y elaborada por el Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial. Una propuesta que no deja lugar a dudas sobre el papel que se espera de la ética: "nuestra visión consiste en establecer la ética como pilar fundamental para garantizar y expandir una IA confiable" (par.13). Como ya hemos mencionado, la experiencia nos previene ante el peligro de que estas directrices queden solo en palabras, que no sean capaces de aplicarse, con rigor y efectividad, a una revolución digital que ya lleva tiempo en marcha. Se habla mucho, pero se hace poco o nada. Con lo que el resultado no es la confianza sino al contrario, la suspicacia y la desconfianza. Hablamos de *confianza*, cuando creemos tener razones para esperar algo, cuando pensamos que tal o cual expectativa va a cumplirse. La *fiabilidad*, por su parte, nos lleva a la probabilidad del buen funcionamiento de algo. Ambas remiten a una *confiabilidad* como capacidad de las personas e instituciones de ser dignas de confianza. La voluntad de una parte de depender de otra no es casual, ni se basa en

intuiciones arbitrarias y ajenas a la experiencia. Confiar no es un sentimiento o una creencia irracional, más bien son las razones las que nos disponen a confiar, las que nos dan ánimo, empuje y aliento para actuar, nos motivan a la acción (García-Marzá, 2004). La Estrategia Europea y su Plan Coordinado dejan bien claro que: "la confianza es un requisito previo para garantizar una IA centrada en el ser humano". La Inteligencia Artificial, nos recuerda la Comisión, no es un fin en sí mismo, sino un medio que debe servir para "mejorar" la vida de las personas. Estas son las expectativas que tenemos y en las que la práctica digital apoya su legitimidad y su credibilidad social.

El marco normativo desarrollado en este documento coincide con los niveles anteriormente descritos que deben recorrer toda ética aplicada para pasar de la validez moral a la práctica, de los principios éticos a la realidad. En suma, para la realización del "deber ser". Ninguno de estos tres niveles puede obviarse porque cada uno de ellos implica una normatividad distinta. El documento que venimos analizando recoge estos tres niveles:

1. Los *fundamentos* de una IA fiable se encuentran en la declaración de derechos humanos y en la Carta Europea de Derechos Humanos. Son principios éticos basados en los derechos fundamentales y, en último lugar, en el reconocimiento de la igual dignidad de todas las personas. Son los siguientes: *autonomía* (respeto de la dignidad humana y reconocimiento del otro como interlocutor válido); *no maleficiencia (*prevención del daño y protección de la dignidad*)*; *justicia* (inclusión y distribución justa de costes y beneficios) y *explicabilidad* (transparencia, información y participación en la toma de decisiones). La finalidad de estos principios éticos es inspirar y guiar la lógica del desarrollo, utilización y aplicación de los sistemas de IA y que, por tanto, definen su responsabilidad, aquello de lo que deben dar razón ante la sociedad.
2. La *realización* de una IA fiable requiere siete requisitos básicos que definen las condiciones para su ejecución a lo largo del ciclo de vida de los sistemas de IA: acción y supervisión humanas; solidez técnica y seguridad; gestión de la privacidad y datos; transparencia; diversidad, no discriminación y equidad; bienestar social y ambiental y rendición de cuentas.
3. La *aplicación* concreta requiere a su vez de un análisis de las condiciones y características propias de los contextos específicos, de la *evaluación* de las posibilidades reales de acción.

No es posible entrar en este artículo en los entresijos de esta arquitectónica. Solo afirmar que nos encontramos ante unas directrices donde, por primera vez en las directrices europeas, la ética juega un papel central tanto en la definición como en la gobernanza de una Inteligencia Artificial centrada en el ser humano. La Comisión es consciente de que la confianza buscada depende tanto del cumplimiento legal como de las condiciones de una tecnología robusta. Pero también, y en especial, depende del consentimiento, voluntario e informado, de todos los agentes y procesos que forman parte del contexto socio-técnico, del acuerdo de todos los agentes individuales e institucionales que participan en la generación y en su gestión. Textualmente:

> Para hacer una IA fiable es preciso garantizar la inclusión y la diversidad a lo largo de todo el ciclo de vida de los sistemas de IA. Hay que tener en cuenta a todos los afectados y garantizar su participación en todo el proceso, también es necesario garantizar la igualdad de acceso mediante procesos de diseño inclusivos, sin olvidar la igualdad de trato (par.79).

Esta participación viene exigida desde el principio de autonomía como reconocimiento de la igual dignidad de todas las personas. Una ética digital dialógica tiene precisamente en esta participación libre e igual, en la deliberación y búsqueda de acuerdos, el pilar central para la generación de confianza. Autores como Karl O. Apel y Jürgen Habermas han desarrollado las bases de una ética discursiva, mientras que en la actualidad se desarrollan, en una segunda ola, las éticas aplicadas derivadas de esta exigencia moral de la participación como requisito para el desarrollo de la autonomía y como sostén del valor intrínseco de la dignidad (Cortina; Conill; García.Marzá, 2008). El fundamento moral de la necesidad de este diálogo y posible acuerdo se encuentra en el reconocimiento del valor intrínseco de dignidad de todas las personas implicadas en la realidad digital. Desde este horizonte de actuación, el documento falla al no concretar las posibilidades y procedimientos para esta participación. Como veremos a continuación, encontramos lagunas en el texto que limitan la confianza que pretenden generar.

El valor moral de las directrices que estamos analizando y, con él, su fuerza vinculante, su obligación y su capacidad de convertirse en recursos morales disponibles por todas las partes implicadas, no desaparece en el segundo y tercer nivel. Más bien en estos niveles de adecuación y concreción los principios éticos del primer nivel deben integrarse con las posibilidades de realización, con los límites y potencialidades que cada realidad concreta nos ofrece. La moralidad, y por lo tanto, la exigibilidad de las decisiones, acciones o instituciones, no desaparece con la aplicación de los principios, como bien muestra nuestra capacidad de valorar moralmente los resultados alcanzados y en los que, como veremos en el siguiente apartado, se apoyan las bases éticas de la confianza.

Sin embargo, en el documento europeo no se aprecia bien *esta trazabilidad moral*. En mi opinión, tres cuestiones básicas deberían replantearse para poder explicitar y gestionar estas directrices éticas para una IA confiable.

En primer lugar, es evidente que la fundamentación de estos cuatro principios no puede estar en la recopilación y en la mayor o menor coherencia entre las actuales directrices internacionales, como afirma Floridi (2018: 696). La validez moral, y, por lo tanto, su obligatoriedad, no dependen de un análisis empírico, de una mera comparación entre diferentes propuestas actualmente existentes. Antes bien, las directrices coinciden porque son conceptos que recogen las experiencias históricas de la protección de la dignidad humana, concretadas en los Derechos Humanos y en la Carta Europea. Los principios éticos descritos en el primer nivel pueden considerarse como condiciones de posibilidad de la realización de la dignidad humana en la esfera digital y en sus prácticas y procedimientos.

En segundo lugar, la semejanza con los principios de la bioética no deriva solo del hecho que la bioética es la que más se parece a la ética digital al "tratar ecológicamente con nuevas formas de agentes, pacientes y entornos" (Floridi, 2019). Más bien habría que insistir en la profunda, y muchas veces insondable, asimetría de poder como rasgo básico compartido

en ambos ámbitos. Asimetría que se da entre aquellos que tienen la capacidad de definir y gestionar los datos y quienes van a sufrir las consecuencias de su conversión en algoritmos, en una nueva realidad. Al igual que en su momento ocurrió con la investigación con seres humanos, los cuatro principios bioéticos se adaptan bien a los nuevos retos éticos que plantea la inteligencia artificial, pues también aquí se trata de convertir a los pacientes en agentes, en ciudadanos digitales. Lo que, a mi juicio, no es comprensible es la razón por la que desparece el principio de beneficencia cuando, de hecho, las palabras "mejora"," bienestar", "calidad de vida", etc. brotan por doquier en el documento.

En tercer lugar, a lo largo del documento la participación se reduce a la información, transparencia y, a lo sumo, monitorización. Se trata, generalmente de una relación unidireccional. Cuando se habla de acción humana se entiende que "los usuarios deberían ser capaces de tomar decisiones autónomas con conocimiento de causa en relación con los sistemas de IA. Se les deberían proporcionar los conocimientos y herramientas necesarios para comprender los sistemas de IA e interactuar con ellos de manera satisfactoria y, siempre que resulte posible, permitírseles evaluar por sí mismos o cuestionar el sistema. Los sistemas de IA deberían ayudar a las personas a tomar mejores decisiones y con mayor conocimiento de causa de conformidad con sus objetivos" (par.64).

La tesis principal del presente artículo se centra en argumentar que el principio de explicabilidad puede responder a gran parte de estas preguntas, rellenar estos vacíos, siempre y cuando no pierda su carácter moral, esto es, no se reduzca a ser una mera estrategia ante el pragmatismo de lo realmente posible (Apel, 1988).

La razón de añadir un nuevo principio a los ya utilizados por la bioética no debemos buscarla solo en la complejidad de los macrodatos, del internet de las cosas o de los algoritmos, ni en la desproporción entre quienes construyen los algoritmos y las máquinas de decisión y aprendizaje y los que van a sufrir las consecuencias. Esta asimetría no nos lleva solo a la necesidad de comprender y rendir cuentas de los procesos de toma de decisiones de la IA, como algunos autores creen (Floridi et. al. 2018: 699), sino también a restituir la falta de reciprocidad y de reconocimiento recíproco a través de la *participación libre e igual y, con ella, la posibilidad de influir de forma efectiva, de discutir las posibilidades de acción, de limitar aquellos desarrollos tecnológicos que no tengan claras las consecuencias según el principio bien conocido de precaución*. Una necesidad que no es meramente estratégica, que no deriva solo de la complejidad, opacidad e ininteligibilidad de las prácticas digitales, sino de la moralidad de un principio que nos permite el paso de lo deseable a lo posible, pues es capaz de vincular internamente principios, requisitos y contextos.

En el documento que venimos analizando se presenta el principio de explicabilidad como "crucial" para que los usuarios confíen en los sistemas de IA y para mantener esta confianza. Esto significa:

> (…) que los procesos han de ser transparentes, que es preciso comunicar abiertamente las capacidades y la finalidad de los sistemas de IA y que las decisiones deben poder explicarse —en la medida de lo posible— a las partes que se vean afectadas por ellas de manera directa o indirecta. Sin esta información, no es posible impugnar adecuadamente una decisión (par.53).

En las diferentes directrices internacionales actualmente existentes este principio se expresa en diferentes términos, con mayor o menor identidad semántica: transparencia, responsabilidad, inteligibilidad, comprensibilidad, trazabilidad, comunicabilidad, apertura, claridad, etc. (Jobin, 2019; Mittelstad et. al., 2016). El principio abarca el conocimiento de cómo funciona el algoritmo y también quién es el responsable de su funcionamiento. De esta forma, este nuevo principio se convierte en el hilo con el que se cose la coherencia con el resto de principios, pues se trata que el ciudadano conozca y comprenda la realidad digital en la que está inmerso (Floridi; Cowls, 2019: 12).

Sin embargo, de la definición del principio de explicabilidad que encontramos en las Directrices ni se sigue, ni se exige, la participación de todos los afectados en la que se apoya su valor moral. Parece más bien que estemos ante un mero proceso informativo y, por lo tanto, vertical. Leemos sobre transparencia, sobre explicación y rendición de cuentas de las decisiones tomadas, sobre quién y cómo ha decidido actuar, etc. Características totalmente necesarias, pero totalmente insuficientes, pues no exigen la presencia y el acuerdo de todas las partes implicadas. Podemos pensar que mejor es poco que nada, pero de nuevo lo que está en duda es el valor moral del principio de explicabilidad y, por tanto, su fuerza. Este aspecto sale a la luz solo con seguir leyendo el documento de las directrices:

> No siempre resulta posible explicar por qué un modelo ha generado un resultado o una decisión particular (ni qué combinación de factores contribuyeron a ello). Esos casos, que se denominan algoritmos de «caja negra», requieren especial atención. En tales circunstancias, puede ser necesario adoptar otras medidas relacionadas con la explicabilidad (por ejemplo, la trazabilidad, la auditabilidad y la comunicación transparente sobre las prestaciones del sistema), siempre y cuando el sistema en su conjunto respete los derechos fundamentales (par.53).

Por lo visto, y así se afirma en el texto, el grado de necesidad de explicabilidad depende en gran medida del contexto y la gravedad de las consecuencias derivadas de un resultado eerróneo o inadecuado. Pero, en mi opinión, tal afirmación no se sostiene desde el punto de vista moral puesto que nada sabemos sobre quién decide las consecuencias posibles en el momento de la recogida de datos o sobre quién define los datos que alimentarán al sistema, por poner dos ejemplos. Como bien sabemos, la autonomía depende de la posibilidad de participar en todo aquello que nos afecta. En esa capacidad se apoya nuestra dignidad y el valor moral de este principio. Valor moral que desaparece si para nada se tiene en cuenta la exigencia de una inclusión libre e igual de todos los posibles afectados. Según el texto anterior, el grado de necesidad de explicabilidad depende en gran medida del contexto y la gravedad de las consecuencias derivadas de un resultado erróneo o inadecuado. ¿Quién decide si el contexto es adecuado o no? ¿quién la gravedad de las consecuencias? Sin una respuesta, los tres principios éticos anteriores al de explicabilidad caen fácilmente en una especie de "justifica-lo-todo".

En mi opinión, la trazabilidad y la auditabilidad, al igual que la claridad y la inteligibilidad, forman parte efectivamente de todo proceso de transparencia. Pero la transparencia es solo la mitad de la explicabilidad. Hace falta la participación que nos permita deliberar acerca de las preguntas que hemos ido dejando sin respuesta, que nos posibilite, por ejemplo,

alcanzar acuerdos acerca de la conveniencia o no de soportar ciertas desventajas. Sin la aceptación libre y voluntaria, no hay autonomía. Sin autonomía no tenemos ninguna base moral para apoyar el reconocimiento de la igual dignidad de las personas. Este es el fundamento moral del principio de explicabilidad y su potencialidad para el diseño institucional del que a continuación nos ocuparemos.

## 3. Del principio kantiano de publicidad al principio de explicabilidad. La aportación al diseño institucional

A primera vista puede parecer una ingenuidad exigir la inclusión de *todos* los posibles implicados en las decisiones que les afecten cuando hablamos de miles de millones de personas, de procesos complejos, muchas veces incomprensibles, de consecuencias que nadie esperaba o preveía, etc. Pero tal candidez desaparece cuando recordamos los tres pasos necesarios que van de la exigencia moral a la realidad, del deber ser de los principios éticos a la pretensión de justicia de cada situación concreta, a la realidad de los contextos sociotécnicos, sociales y económicos.

De ahí que debamos mostrar que la validez moral de los principios éticos presentados, la exigencia de universalidad y, por lo tanto, su valor de convicción, su fuerza, sigue estando cuando descendemos al nivel de realización, a los requisitos necesarios para posibilitar y garantizar su aplicación, siempre mediada a través de las condiciones empíricas de realización. Este es, a nuestro juicio, el papel clave que juega el principio de explicabilidad, pues nos permite hablar de una *estrategia moral*, de una mejor o peor aproximación a la idea de la participación y posible acuerdo de todas las partes implicadas (Apel, 1988; García-Marzá, 2019). En mi opinión, una comparación del principio de explicabilidad con el *principio de publicidad* introducido por Kant en la *Paz Perpetua* nos permitirá entender este funcionamiento de la ética digital que, como ética aplicada, debe dar razón del "deber ser" incrustado en la realidad de toda práctica digital.

Con el *principio de publicidad* Kant se propone mediar entre los principios éticos y la práctica, entre lo deseable y lo posible, sin que esta mediación pierda su valor moral (García-Marzá, 2012). Quizás refrescar algunas cuestiones de su propuesta nos sirva para aclarar el valor moral del principio de explicabilidad y la exigencia de participación qu7e conlleva. Cuando Kant nos avisa del peligro que encierran aquellas instituciones donde "todos sin ser todos deciden" nos previene de las consecuencias de anular la diferencia entre la exigencia moral, el segundo *todos* de la frase, y su concreción práctica, el primer todos. Esta distancia entre los principios éticos y su posible realización práctica nunca puede recorrerse del todo, por más justificada que esté la estructura participativa de la empresa, la universidad, el parlamento, el laboratorio, etc. Es precisamente el recorrido, más o menos largo, entre el "todos moral" y el "todos fáctico" donde se asienta toda perspectiva crítica y desde donde se construyen las razones que apoyan la confianza (García-Marzá, 2004).

Esta es la idea básica que desarrollará Kant con el principio de publicidad, cuya primera definición nos habla de la injusticia o inmoralidad de las decisiones tomadas y es bien sencillo: *"Todas las acciones referidas al derecho de otros hombres cuya máxima no es compatible con la publicidad, son injustas"* (ZeF, VIII, 381). Aplicado a nuestro campo, si

los algoritmos deben ser secretos es, de forma clara y tajante, porque son injustos. Podemos buscar eximentes en los derechos de propiedad o en las patentes, pero la validez moral es proporcionalmente inversa a la opacidad. Si nuestra decisión o acción, si los presupuestos, por ejemplo, el origen de los datos, no pueden hacerse públicos es porque son injustos. Desde este principio, todo elemento de los sistemas de Inteligencia Artificial que no pueda ni comunicarse, ni explicarse, es inmoral. Esto significa que deben descartarse. Los algoritmos secretos no pueden ser éticos.

Pero a continuación Kant también nos ofrece una definición positiva del principio de publicidad, aunque ya no tan tajante como la negativa. Dice así: *"Todas las máximas que necesitan de la publicidad (para no fracasar en sus propósitos) concuerdan con el derecho y la política a la vez"* (*ZeF,* VIII, 386).

Este segundo principio ya no exige sólo hacer visibles y conocidas las razones, decisiones, procedimientos, etc., sino que se refiere también a una especie de unión o coincidencia de todos los afectados: la posibilidad de un conocimiento público conlleva, por así decirlo, la necesidad de la aceptación o aprobación de los demás, de otra forma podría fracasar en sus propósitos. De su consentimiento, en definitiva. Aquí público adquiere un segundo significado: ya no se opone sólo a secreto, también parece oponerse a cerrado pues requiere de la aprobación del público para que se cumpla. Necesitar de la publicidad significa que los algoritmos, y demás elementos de los sistemas de IA, no podrían tener éxito, ser eficaces, sin que fuesen públicos. Sencillamente por la desconfianza que generan.

Por supuesto, este acuerdo posible no puede ser anticipado por una o por varias de las partes implicadas, por ejemplo, programadores, empresas, agencias gubernamentales, etc. Si bien no podemos anticipar el acuerdo, sí que existen mecanismos para saber si nos alejamos o nos acercamos a la idea del consenso posible de todos los implicados. Al final de la obra, Kant nos propone un paso más en esta relación entre los principios y su realización, en la explicitación del potencial crítico que encierra el principio de publicidad (García-Marzá, 2012). Como si de una tercera formulación se tratara, Kant equipara la publicidad con la eliminación de toda desconfianza. Textualmente: "Si sólo mediante la publicidad puede lograrse este fin, es decir, mediante la eliminación de toda desconfianza respecto a las máximas, éstas tienen que estar en concordancia con el derecho del público, pues sólo en el derecho es posible la unión de los fines de todos". (ZeF, VIII, 386) ¿No es esta la confianza que buscan las directrices éticas?

La justificación moral y no solo instrumental del principio de explicabilidad deriva, por tanto, del reconocimiento y respeto de la autonomía de todos aquellos afectados o implicados por la regulación o legislación. Kant remite esta justificación a la razón pública, definida como una facultad "donde todos tienen voz". En esta participación de todos los implicados y afectados, personas e instituciones, se basa la justificación moral del principio de explicabilidad. Las tres formulaciones derivan del reconocimiento recíproco de todos aquellos implicados en la regulación institucional, del carácter insustituible de la voluntad libre en la que se asienta la dignidad de las personas. Esta voluntad es la que exige nuestro consentimiento o acuerdo, nuestra libre aceptación. Solo sobre esta posible conformidad es posible generar y garantizar la confianza. Desde esta justificación moral, la aplicación del principio de explicabilidad, entendido como la suma de transparencia y participación *siempre* debe ser posible.

Sin este principio, capaz de mediar entre el *todos moral* y el *todos pragmático,* el resto de principios deja de tener sentido y nos quedamos sin un criterio de lo que es correcto o incorrecto. Un criterio que debe aplicarse, como a continuación veremos, desde el inicio, desde la definición de los datos y el diseño de los algoritmos. En esta tensión inherente al principio de explicabilidad queda integrada la *responsabilidad*, la capacidad de dar razones de lo que hacemos o dejamos de hacer ante los afectados. Más aún, la responsabilidad se convierte siempre en corresponsabilidad. De ahí la generación de confianza.

Veamos cómo se desarrolla este principio de explicabilidad en los requisitos que, según el documento que comentamos, la aplicación de una IA fiable exige. Solo cuando entramos en el terreno de los requisitos o condiciones para la aplicación de los principios nos encontramos con la acción y la supervisión humanas dividida a su vez en tres niveles:

1) *Participación humana*: capacidad de que intervengan los seres humanos en todos los ciclos de decisión del sistema, algo que en muchos casos no es posible ni deseable.
2) *Control humano*: capacidad de que intervengan seres humanos durante el ciclo de diseño del sistema y en el seguimiento de su funcionamiento.
3) *Mando humano*: capacidad de supervisar la actividad global del sistema de IA, incluidos sus efectos económicos, social, jurídicos y éticos, así como la capacidad de decidir cuándo y cómo utilizar el sistema en una situación determinada (par. 65)

Sin embargo, a la hora de concretar estas exigencias morales derivadas de nuestra autonomía, se afirma rotundamente que la intervención de todos los seres humanos *"no es posible ni deseable*". Lo primero es evidente, lo segundo es una afirmación que rompe con la misma justificación moral que se pretende. De esta forma, todo el sistema de una Inteligencia Artificial confiable pierde valor moral y, por lo tanto, fuerza y eficacia. La participación pasa de considerarse condición de posibilidad de la confianza, de ser moralmente exigible, a ser solo "recomendable" consultar a las partes interesadas que pueden ser afectadas directa o indirectamente por el sistema. Más aún, si la participación de todos los afectados resulta, incluso, "indeseable". ¿Cómo confiar en un sistema que no depende de nosotros? ¿Cómo gestionar, incluso conocer, los límites de la tecnología? ¿Cómo definir lo posible y lo imposible?

El paso del "*todos"* reflejado en los principios éticos (nivel 1) al "*todos"* pragmático en las diferentes prácticas e instituciones (niveles 2 y 3) es una de las dimensiones más importantes de la reflexión ética y la clave para convertir el saber moral en un recurso moral, finalidad última de toda ética aplicada (García-Marzá, 2004). Solo así nos acercaremos a las bases éticas de la confianza en la Inteligencia Artificial.

Para esta motivación, para generar confianza, no es suficiente con una declaración de buenas intenciones por parte de los profesionales o de sus organizaciones. Desde el principio de explicabilidad como principio ético toda gestión de la información que pretenda validez moral debe pasar, en cada situación concreta, por hacer públicos los esfuerzos realizados. No se trata sólo de una disposición a la sinceridad, sino de que esta disposición adquiera el rango de un compromiso público, en el doble sentido de trasparencia y de participación que ya hemos analizado. Con esta idea trabajan las teorías del diseño institucional al remitir la capacidad de producir confianza a este "potencial de justificación discursiva" (Goodin, 2008).

Diseñar parece un término pretencioso y arriesgado, pero esta primera impresión desaparece cuando nos percatamos que su raíz etimológica designare nos indica la tarea de señalar qué institucionalización d ellos requisitos es la más adecuada a un contexto particular. Si bien diseñar o rediseñar son actividades intencionales, deben entenderse siempre como aportaciones a una deliberación pública acerca de qué infraestructura ética es la más adecuada para que nuestras organizaciones generen confianza. Es decir, acerca de cómo sostener y desarrollar la credibilidad y la reputación de nuestra organización. Dicho de otro modo, para responder de esta justificación pública no basta con la buena voluntad del profesional, sino que debemos contar con procesos y estructuras organizativas que permitan y potencien las directrices éticas señaladas.

Para anclar estas bases éticas de la confianza necesitamos tanto la transparencia, trazabilidad e inteligibilidad de la información, como la posibilidad de que los grupos de interés o sus representantes puedan participar desde la declaración de utilidad, hasta el cálculo de resultados, pasando por el mismo diseño. No hay transparencia sin posibilidad de participación, sin poder decidir, por ejemplo, de qué se informa o cómo calculamos las consecuencias y para quién son. Por así decirlo, el riesgo moral, la posibilidad de que otros sufran las consecuencias de mis decisiones o prácticas, de que no ocurra lo que esperábamos, es directamente proporcional a la participación. No hay autonomía sin posibilidad de ser incluido en las decisiones que acabarán afectándonos. Y esta participación, como muy bien resalta el documento, en todo el ciclo de vida del algoritmo, por supuesto también en su creación. No solo debemos acompañar a la tecnología, debemos adelantarnos, ir siempre un paso por delante (Etzioni 2017; Dignum, 2018).

De acuerdo con estas exigencias, esta es *la propuesta de una ética digital dialógica*, un diseño capaz de responder y de facilitar la aplicación del principio de explicabilidad, un diseño que no abandone las decisiones en una expertocracia irresponsable, debería adquirir la forma de una infraestructura ética con cuatro elementos básicos, adaptables a cada situación particular y a cada estructura organizativa concreta:

1. *Códigos ético y de conducta* El primer paso en esta generación de confianza lo constituye la elaboración y publicación de los códigos éticos y de conducta. Se trata de documentos formales donde encontramos una declaración explícita de los valores que deben orientar la conducta de empleados y directivos, propiciando así las buenas prácticas y marcando el carácter y la personalidad de la organización. Su función es, por lo tanto, doble: - desde el punto de vista interno, formalizar los valores y criterios de decisión que definen la cultura organizativa; desde el punto de vista externo, gestionar la reputación de la organización. No sólo nos presenta los valores que definen el carácter o ética de la organización, sino también los compromisos que está dispuesta a asumir para crear esta voluntad común y las conductas necesarias para su realización. Unos códigos que deben incluir en su seno su compromiso con los sistemas internos de cumplimiento y con las auditorías externas (García-Marzá, 2017).
2. *Comité de ética*. Se concibe como un *espacio de participación* y diálogo de los diferentes grupos de interés en el interior mismo de la organización, encargado del seguimiento y control del programa de ética y cumplimiento, así como del impulso

de las directrices éticas y sus diferentes procedimientos. Su función es triple: asesorar en temas relacionados con la interpretación y aplicación del código ético; resolver las notificaciones de sugerencias, alertas y denuncias realizadas a través de la línea ética; promover la información y formación de los empleados y directivos en el programa de ética y cumplimiento. La confianza en el comité dependerá, a su vez, de la confianza que sean capaces de generar sus componentes, como muy bien ha mostrado el "fiasco" del comité de ética de Google.

3. *Línea ética:* la participación buscada no puede limitarse a un pequeño comité que, aunque aporte la presencia y voz de los grupos de interés internos y externos, no sustituye a la voz de todos. Debemos establecer canales de comunicación que permitan la participación de todo aquel que quiera hacerlo, siempre centrada en el cumplimiento de los compromisos éticos adoptados. La comunicación no debe limitarse a la denuncia de malas prácticas. También, y en especial, debe potenciar una cultura ética a través de la implicación de los empleados en la formación y el desarrollo, en la gestión, en suma, de los valores éticos. Esta participación debe incluir a los grupos *internos* (investigadores, desarrolladores, directivos, trabajadores, etc.), como *externos* (compañías de la competencia, agencias gubernamentales, consumidores, organizaciones de la sociedad civil, etc.

Con estos tres instrumentos de gestión de la ética, estamos ante diferentes pasos progresivos para la generación de confianza en todo el tejido socio-técnico de la realidad digital. Ahora bien, la existencia y el funcionamiento de esta infraestructura ética tiene, a su vez, que ser verificada externamente. Este es el papel de la auditoría ética, de la evaluación de la IA fiable que nos proporciona el documento (par.112).

4. La *auditoría ética*. La auditabilidad se refiere en el documento a la capacidad de un sistema de IA de someterse a la evaluación de sus algoritmos, datos y procesos de diseño. De ahí que forme un elemento fundamental para el seguimiento de la participación (Buruk et. al, 2020).

Como conclusión, al diferenciar claramente entre tres dimensiones básicas para una IA confiable: legal, ética y robusta, las directrices se refieren también, por consiguiente, al *riesgo moral*, a los riesgos derivados del incumplimiento de los principios éticos y de sus requisitos de aplicación. No se trata de sancionar, sino de crear una cultura donde la transparencia y la participación dificulten las malas prácticas y reconozcan y potencien las buenas. La gestión de la confianza es inversamente proporcional a este riesgo moral, a la desconfianza que produce no saber si la organización va a cumplir o no con lo que se espera de ella. Este es el principal objetivo del principio de explicabilidad y la justificación moral de la exigencia de participación que le es inherente.

**Referencias**

Apel, K.O., (1988). *Diskurs und Verantwortung das Problem des Übergangs zur postkonventionellen Moral,* Frankfurt: Suhrkamp

Buruk, B., Ekmekci, P. E., & Arda, B. (2020). A critical perspective on guidelines for responsible and trustworthy artificial intelligence. *Medicine, Health Care and Philosophy*, *23*(3), 387-399.

Calvo, P (2021). El gobierno ético de los datos masivos. *Dilemata. Revista internacional de* éticas aplicadas, (34), 31-49

Cortina, A., Conill,J; García-Marzá (eds.) (2008). *Public reason and applied ethics: The ways of practical reason in a pluralist society*. Londres: Ashgate Publishing.

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics Inf Technol* (20), 1–3.

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, *21*(4), 403-418.

European Data Protection Supervisor (2015). *Hacia una nueva ética digital. Datos, dignidad y tecnología* (Dictamen 4/2015). ESPD. Recuperado de: https://edps.europa.eu/sites/edp/files/publication/15-09-11_data_ethics_es.pdf.

COM(2019) 168 final (2019). *Generar confianza en la inteligencia artificial centrada en el ser humano*. Bruselas: Comisión Europea.

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

Floridi, L. (2018). Soft ethics and the governance of the digital. *Philosophy & Technology*, *31*(1), 1-8.

Floridi, L. Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689-707.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, *32*(2), 185-193.

García-Marzá, D. (2004). Ética empresarial: del dialogo a la confian*z*a. Madrid: Trotta

García-Marzá, D. (2012). Kant´s Principle of Publicity, *Kant-Studien. Philosophische Zeitschrift der Kant-Gesellsschaft*, (103), 96-113.

García-Marzá, D. (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, D. (2019). Repensar la democracia. Estrategia moral y perspectiva crítica en KO Apel. *Daimon Revista Internacional de Filosofía*, (78), 75-89.

García-Marzá, D. & Calvo, P. (2022). Democracia algorítmica: ¿un nuevo cambio estructural de la opinión pública?. *Isegoría*, (67), e17-e17, 1-16. https://doi.org/10.3989/isegoria.2022.67.17

Generalitat Valenciana (2018). *Estrategia de Inteligencia Artificial de la Comunitat Valenciana*, Valencia: GVA.

Goodin, R. E. (Ed.) (1998). *The theory of institutional design*. Cambridge: Cambridge University Press.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Kant, I. (1795). *Zum ewigen Frieden. Ein philosophischer Entwurf*, (ZeF), AA, VIII.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 205395171667967. https://doi.org/10.1177/2053951716679679

Rehg, W. (2015). Discourse ethics for computer ethics: a heuristic for engaged dialogical reflection. *Ethics and Information Technology*, *17*, 27-39.

UNESCO (2022). *Recomendación sobre la ética de la Inteligencia ArtificiaL de UNESCO*. Montevideo, Uruguay: UNESCO.

Yuste, R., Goering, S., Arcas, B. A. Y., Bi, G., Carmena, J. M., Carter, A., Fins, J. J., Friesen, P., Gallant, J. L., Huggins, J. E., Illes, J., Kellmeyer, P., Klein, E., Marblestone, A. H., Mitchell, C., Parens, E., Pham, M. Q., Rubel, A., Sadato, N., . . . Wolpaw, J. R. (2017). Four ethical priorities for neurotechnologies and AI. *Nature*, 551(7679), 159-163. https://doi.org/10.1038/551159a