APROXIMACIÓN A LA ÉTICA DIGITAL
HACIA UN DESARROLLO SOCIALMENTE PREFERIBLE DE LOS SISTEMAS DE INTELIGENCIA ARTIFICIAL

CAMILO GUZMÁN VELÁSQUEZ

JUAN GUILLERMO LALINDE

Tesis de grado

Asesor:

Juan Guillermo Lalinde

UNIVERSIDAD EAFIT

Escuela de Ingeniería

Maestría en Ingeniería - Modalidad Investigación

Medellín - Colombia

2021

Abstract

La 4ta Revolución ha implicado reflexionar sobre los efectos sociales, económicos y ambientales de las tecnologías digitales. Especialmente en la última década, la Inteligencia Artificial se ha convertido en uno de los temas centrales de esta reflexión, identificando los efectos que estas poderosas tecnología tienen y tendrán, en cada uno de los niveles del desarrollo de la vida humana. Por lo tanto, se requiere de un esclarecimiento conceptual que guíe hacia el entendimiento de los fenómenos y las transformaciones propias de una revolución. Así, el presente artículo analiza el reciente y amplio término *ética digital* como posible marco de referencia a la hora de abordar dilemas éticos, y que a su vez, procura cosechar el potencial de estas tecnologías. Para esto, **1)** Se argumentará en torno al rol determinante de los datos y algoritmos en las actuales sociedades digitales y de la información. **2)** se definirá el concepto *ética digital,* su naturaleza y alcance. **3)** Para posteriormente profundizar en una de sus ramas, la *ética de la Inteligencia Artificial,* ubicando en ésta los *principios filosóficos* que la sustentan. **4)** Luego, en el marco de este concepto, se analizará la figura *Ethics as a service*, como un dispositivo del contexto organizacional que habilita la implementación práctica de los principios éticos en el proceso de diseño y desarrollo de los algoritmos y servicios basados en IA. **5)** Y por último se presentará la *auditoría basada en la ética*, como un mecanismo de gobierno de la inteligencia artificial que procura revisar y comprobar constantemente la apropiada materialización de los principios éticos en los algoritmos y servicios basados en IA. Todo lo anterior con el propósito de diseñar tecnologías con efectos positivos y reducir las consecuencias negativas para la sociedad.

Palabras clave

Datos, Inteligencia Artificial, Ética Digital, Ética de la Inteligencia Artificial, Auditoría basada en la ética, Ética como servicio

1 Introducción

Toda disciplina científica dispone y genera un discurso que le habilita hacer frente a los retos y problemáticas propias del área. La biología, la geología, la psicología, entre muchas otras, tienen sus propios conceptos, teorías y métodos, que permiten la investigación y comprensión de los fenómenos de los que se ocupan. La especialización de los discursos y disciplinas ha llevado, en cierta medida, a la fragmentación del conocimiento y a la falta de una comprensión de las interrelaciones que hay entre las distintas disciplinas científicas (Morin, 1999). Pero, pareciera que la historia cada vez nos recuerda con más insistencia que un estilo de pensamiento fragmentado no será la forma más apropiada de aproximarnos a los retos de este siglo, definido por la revolución digital (Villoro, 2020).

Así, podríamos decir que la revolución digital ha incidido especialmente en la manera en que entendemos ciertos conceptos esenciales para la vida, como lo son el tiempo, el espacio y la naturaleza humana (Floridi, 2007). Esto quiere decir que las herramientas conceptuales que traíamos ya no son suficientes para dar explicación y sentido a las nuevas dinámicas y realidades digitales (Taddeo, 2016). Por lo tanto, las ciencias y en especial la filosofía adquieren una nueva tarea, la de diseñar conceptos que logren abarcar los nuevos entornos que las sociedades de la información implican (Floridi, 2015). Es por esto, que en medio de la llamada revolución digital, y con todo el caos que puede causar a la hora de querer comprender sus efectos, **el presente artículo pretende hacer un ejercicio de comprensión,** frente a una problemática que poco a poco está incidiendo en cada una de las esferas de la vida humana y de las disciplinas científicas, **la carencia de puentes entre las ingenierías, la ciencia y la ética** (De Cremer & Kasparov, 2021)

El tiempo en que hoy vivimos, está caracterizado por cómo las tecnologías de la información y la comunicación ya no solo están relacionadas a nuestro bienestar individual y social, sino que nuestro bienestar y nuestras sociedades están soportados en estas tecnologías (Floridi, 2011). Así, gracias a ellas y en especial al internet, las actuales sociedades de la información han podido erigirse y trascender el espacio físico y geográfico que las delimitaba, trayendo retos sin precedentes (Floridi, 2012). Es debido a lo digital que se abren nuevas fronteras de interacción y organización social, ahora incluyendo a los agentes de inteligencia artificial, lo cual nos lleva a replantearnos nuestro rol como seres humanos en el entorno digital que habitamos (Taddeo, 2016).

Esta, se convierte entonces en una época en la que las humanidades y las ciencias de la computación comienzan a abrir espacios de conversación para explorar las nuevas formas de observar, intervenir y modificar los comportamientos humanos y los contextos que habitan. Es así, como surgen la ética digital y las ciencias de datos sociales. Dos subdisciplinas, la primera de la filosofía y la segunda de las ciencias del comportamiento, que buscan avanzar en la tarea de esclarecer conceptualmente los fenómenos de la vida digital, y eventualmente, ofrecer herramientas conceptuales para intervenir comportamientos y diseñar contextos.

Y es gracias a estos desarrollos teóricos y conceptuales, que se puede comenzar a vislumbrar el real impacto que tecnologías como la Inteligencia Artificial tiene en los distintos niveles de la sociedad. Siguiendo la línea del trabajo realizado por académicos como Luciano Floridi, Kate Crawford, Sandra Wachter, Christopher Burr, entre otros, se comienza a ampliar el espectro de los impactos que la llamada Inteligencia Artificial tiene en el medio ambiente, en las distintas formas de trabajo necesarias para sostener la casi mágica idea de la IA, en los tipos de datos necesarios, en las implicaciones sociales y políticas, y en la materialidad de estas tecnologías.

Entonces, habiendo enmarcado el contexto en el que actualmente nos encontramos como sociedad, altamente digitalizada, en la que hay más dispositivos digitales conectados a la red que seres humanos, en la que la vida offline se fusiona con la vida online, y en la que los seres humanos interactúan más y más con sistemas de IA sin siquiera darse cuenta (Floridi, 2015); se hace necesario comenzar a esclarecer con mayor detalle los pilares de las sociedades digitales con el fin de diseñar una mejor *Infoesfera* (Floridi, 2014). Para esto, la **sección 2**, ubica el rol de los datos y los algoritmos en el funcionamiento de las sociedades digitales. Siendo los datos la materia prima a partir de la cual se desarrollan productos y servicios. Además, los algoritmos cobran relevancia en este contexto, puesto que la cantidad de datos que las personas, organizaciones y sociedades generan diariamente es abismal (Big Data), por lo tanto se hace necesario una tecnología que pueda ayudar a manejarlos e interpretarlos y así conseguir información de valor. Y con estas premisas avanzaremos hacia una definición más comprehensiva de lo que es la Inteligencia Artificial, a saber: aproximaciones técnicas, prácticas sociales e infraestructuras industriales. Posteriormente, en la **sección 3** abordaremos el reciente término *Ética Digital,* como herramienta conceptual especialmente relevante a la hora de interpretar las problemáticas y retos de las sociedades

digitales, donde los datos, los algoritmos, y las prácticas que los profesionales hacen de estos, pueden ejercer sobre el bienestar de sociedades enteras, y aportar o dificultar a problemáticas de escala global. En la **sección 3.1,** se define y explora el concepto *ética de la inteligencia artificial*, ubicando en éste los *principios éticos* sobre los cuales, gobiernos, entidades multilaterales y academia, de la Unión Europea, Estados Unidos y algunos países de Asia, están acordando dirigir el desarrollo de estas tecnologías. Posteriormente, en la **sección 3.1.2** se aborda el concepto *Ethics as a Service*, como un dispositivo organizacional que procura la materialización del *qué* (principios) y el *cómo* (implementación), esto es, para materializar los *principios éticos* en las etapas de diseño, buscando con esto el desarrollo de una IA *socialmente preferible*. Y en la última **sección 3.1.2,** se sugiere la figura de *la auditoría basada en la ética* como mecanismo de gobierno de la IA la cual busca asegurar que los principios éticos sean efectivamente materializados y auditados en todas las etapas del desarrollo de estas tecnologías.

2 El rol de los datos y los algoritmos en las actuales sociedades digitales

Es bien sabido que dos de los pilares de las sociedades digitales y de la información, son los datos y los algoritmos. Es gracias a estos que las organizaciones y sociedades toman las llamadas "decisiones basadas en datos", buscando mejor pertinencia, precisión e incluso anticipación, en sus acciones (Provost, 2013). Además es gracias a esta infraestructura que las nuevas empresas, o apps, pueden emerger, ofreciendo soluciones convenientes a sus usuarios. Las bien conocidas Uber, Google, Apple, Amazon, Rappi, Facebook, entre muchas otras, son ejemplos de empresas que ofrecen soluciones basadas en infraestructuras digitales, datos e inteligencia artificial. Así, en las siguientes subsecciones profundicemos un poco en lo referido a los datos, las economías digitales, la inteligencia artificial y algunos de los aspectos éticos que suponen.

2.1. Los datos en las economías digitales

Comencemos con una breve definición de lo que son los datos: "son una representación simbólica, ya sea numérica, alfabética, espacial, etc., de un atributo o variable cuantitativa o cualitativa" (Wikipedia, 2021). Así, los datos son descripciones parciales de hechos empíricos, sucesos y entidades. Y solo cuando se tiene un conjunto de datos enmarcados a la

luz de una hipótesis, teoría o pregunta, es cuándo se pueden convertir en información de valor.

Es acá donde comienza a funcionar la llamada economía de la vigilancia propuesta por Shoshana Zuboff (2020). Puesto que en las actuales sociedades digitales, las principales compañías tecnológicas como Google, Facebook, Amazon, Apple, Microsoft, entre muchas otras, recogen una cantidad abismal de datos privados sobre sus usuarios, de tipo individual y grupal, otorgándoles un poder sin precedentes. Es una concentración de poder enorme, pues recolectan datos de forma intensiva sobre las personas, poblaciones y dinámicas de la red, en todo el mundo, sin ninguna regulación (Laux et. al., 2021; Veliz, 2020). Y hay que hacer la puntualización en lo referido a la recolección de datos a nivel individual y grupal, puesto que en la mayoría de los casos, estas compañías no están interesados en analizar los datos a nivel individual; sino especialmente a nivel grupal, esto es, sobre los individuos que comparten características como nacionalidad, raza, género, edad, y más privadas como, rasgos psicológicos, traumas, pérdidas, orientaciones políticas y sexuales, todo con el propósito de influir en sus procesos de tomas de decisiones (Burr, 2019; Floridi, 2013).

El fin último es influenciar las decisiones de los usuarios. Y con un agravante más serio en el caso de las compañías que se hacen llamar “redes sociales” como Facebook y Google, pues venden los insights que han generado de la basta cantidad de datos privados de sus usuarios, al mejor postor en sus servicios de publicidad personalizada. Lo cual está erosionando las estructuras democráticas, además de la autonomía de sus usuarios (Van Bavel, 2021), al fragmentar las narrativas compartidas que cohesionan a las sociedades. Es claro el especial énfasis que se hace sobre la recolección y uso de los datos personales por parte de las compañías digitales en esta época, pues son un tema neurálgico para las sociedades actuales. Pero cuando nos referimos a datos, es en el amplio espectro del concepto. Esto es, datos generados por dispositivos, vehículos, organizaciones, ciudades, monitores del clima, entre muchas otras fuentes, que gracias a sensores de medición sofisticados, es posible recolectarlos y posteriormente analizarlos para tomar decisiones más precisas (Ömhan et. al., 2019). Así, los datos se pueden usar para comprender fenómenos, para predecir posibles ocurrencias, y también, en algunos casos para prescribir aquello que los investigadores determinen.

Teniendo un poco más de claridad respecto a que son los datos, de dónde provienen y para se pueden usar, avancemos en una definición comprehensiva de la inteligencia artificial,

el rol de los algoritmos en esta infraestructura digital, y cómo las organizaciones y sociedades hacen uso de la llamada inteligencia artificial para analizar y darle sentido a la inmensa cantidad de datos (Big Data) que recolectan y usar sus hallazgos para tomar decisiones.

2.2 Inteligencia Artificial, una mirada más allá del paradigma técnico

Los últimos 15 años han traído increíbles avances en el desarrollo de la Inteligencia Artificial, pero a las implicaciones sociales de estas tecnología no se les ha dado la suficiente relevancia (AI Index, 2021; Veliz, 2021), y nos encontramos en un momento histórico de inflexión, en donde el poder y alcance de estas tecnologías está sobrepasando nuestras habilidades para comprender su amplio impacto y cómo funcionan (Bird et al., 2016). Y en la medida en que las tecnologías digitales y la inteligencia artificial inciden en más áreas de nuestras vidas, nos presentan retos sociales, políticos y ambientales que antes eran desconocidos (Sunstein, 2020; Ömhan et al., 2019; Taddeo, 2010).

Ahora, mucho se ha hablado de la IA y cómo ésta podría llegar a "dominar" e incluso "reemplazar" a los seres humanos cuando por fin se alcance la anhelada Inteligencia Artificial General (Kurzweil, 2006), atribuyéndole un alcance omnipotente, y por lo tanto, desresponsabilizando a sus diseñadores y creadores (Watson, 2020). Pero el devenir de los últimos años, ha revelado como esta narrativa esconde intereses particulares, especialmente de compañías radicadas en Silicon Valley (Williams, 2021; Milligan, 2018), y no obedece a la realidad. El hype ha distraído la atención de la opinión pública, de académicos, reguladores, políticos y diseñadores que podrían estar dedicando su energía a proponer formas comprehensivas de cómo articular estas poderosas tecnologías de forma benéfica con la sociedad (Floridi, 2019; Burr, 2020).

Entonces ¿a qué se refiere la idea de Inteligencia Artificial? Abordaremos esta pregunta desde 3 perspectivas que servirán para construir una visión comprehensiva de la IA (Crawford, 2016; Floridi, 2019; Mökander, 2020)

> **Aproximaciones técnicas**: se agrupan todos los avances técnicos como lo son symbolic logic, expert system hasta machine learning. Esta última siendo también una constelación de muchas técnicas como, deep learning y generative adversarial networks. Usar el término inteligente para sistemas como estos, es una trampa pues están muy lejos de ser inteligentes (Crawford, 2021). Estás, realmente son tecnologías muy buenas en **identificar patrones, agrupar, optimizar y hacer**

predicciones sobre grandes cantidades de datos.

Prácticas sociales: Acá se hace referencia a las personas que diseñan estos sistemas, quienes deciden qué problemas se priorizan resolver a través de estas tecnologías. Y este poder, determina crucialmente los tipos de tecnologías que tenemos hoy en día, incluyendo los tipos algoritmos y datos. Pero más importante aún, esta concentración de poder, define qué poblaciones se ven más beneficiadas por estas herramientas y qué poblaciones se ven excluidas e incluso discriminadas (Crawford, 2020; Wachter & Mittelstadt, 2021).

Infraestructuras industriales: AI es una infraestructura enorme que requiere de una red computacional planetaria. Y los recursos tanto económicos como ambientales, necesarios para mantener estas gigantes infraestructuras, son inmensos. Así, estas infraestructuras representan una profunda concentración de poder (Crawford & Schultz, 2019). Por ejemplo, la extracción de minerales raros es una industria que pareciera no tener mucha relación con la inteligencia artificial, pero además de ser indispensable para darle materialidad (hardware) a estas tecnologías, es una industria que genera enormes impactos ambientales y sociales. Y mientras un obrero en Mali que extrae litio para venderlo a empresas tecnológicas como Amazon, puede ganar 1 USD por día de trabajo; Jeff Bezzos, anterior CEO de Amazon, gana aproximadamente 270 millones de USD al día, teniendo en cuenta que ambos trabajan para la misma industria (Crawford, 2017, 2018).

Entonces, preguntarnos por cuáles son los verdaderos costos de estas tecnologías, parece ser una inquietud pertinente y necesaria para esta época. Y teniendo en cuenta lo anterior, queremos argumentar que nos encontramos en una década donde las implicaciones éticas, sociales y ambientales deben abordarse primero o al menos a la par de los asuntos técnicos, pues los impactos de estas tecnologías tienen un alcance global. Y como varios investigadores lo han sugerido ya, el tiempo no está de nuestro lado (Floridi, 2021; Crawdord, 2018). Por lo tanto, un marco conceptual como la *ética digital* se convierte en unos lentes que permiten observar de manera comprehensiva los nuevos dilemas que estas tecnologías traen como lo son la inequidad, la privacidad (Lauer, 2020; Veliz, 2020), la

autonomía, la determinación personal (Van den Babel, 2021), la soberanía digital (Floridi, 2021), entre otros.

3 La ética Digital: Su naturaleza y alcance

Debido a la transformación digital por la que está cruzando la sociedad, la vida individual y la vida colectiva, están teniendo transformaciones profundas en los vínculos, en las formas de trabajo, en la toma de decisiones, en las normas sociales, en la confianza interpersonal e institucional. Así, la reflexión ética se hace cada vez más relevante. Y especialmente para aquellos profesionales y diseñadores de estas tecnologías, pues como ya hemos mencionado, estas tienen un alcance planetario (Floridi, 2021; IEEE, 2019; ACM, 2020).

Teniendo en cuenta que la revolución actual es producto de los avances en las tecnologías digitales, hay un acuerdo común en definir a las actuales sociedades como las sociedades de la información o sociedades digitales (Lai & Viering, 2012). Esto quiere decir que la información es un activo de interés económico, político y social en cada una de las acciones humanas (Floridi 2006, 2011, 2014, 2019; Harari, 2016, 2018). Por eso, preguntas como: ¿Qué características tiene esta información?, ¿Qué datos se almacenan y cómo se usan?, ¿Cómo se construyen las infraestructuras digitales por donde fluye esta información? ¿Cómo los seres humanos son más persuasibles en función de la forma en que se presenta la información?, ¿Quienes son los diseñadores de estas infraestructuras digitales? y ¿sobre qué valores éticos y morales están fundamentadas estas infraestructuras digitales?, son algunas de las preguntas que la ética digital se encarga de abordar.

La conciencia sobre la relevancia, los efectos y las posibilidades que la información y las tecnologías digitales pueden tener en la vida cotidiana apenas comienza a cobrar fuerza en nuestra sociedad. Ampliando así, la brecha entre aquellos que utilizan y sacan provecho del manejo de la información y aquellos que no lo hacen por desconocimiento o por la enorme brecha digital (Floridi, 2014; Harari, 2016). Según la UNESCO, con la emergencia de Internet y las tecnologías de la información y la comunicación (TIC´s), es necesario considerar la alfabetización digital como una de las habilidades indispensables en que es necesario formar para el siglo XXI, tanto niños, jóvenes y adultos (Lai & Viering, 2012); además de formar con fundamentos éticos a los profesionales en ingeniería y ciencias que diseñan las actuales y futuras tecnologías (ACM, 2021).

Así, Carl Ömhan del Oxford Internet Institute (2019) sintetiza elegantemente:

> La revolución digital brinda enormes oportunidades para mejorar la vida pública y privada, y nuestros entornos, desde la atención médica hasta las ciudades inteligentes y el calentamiento global. Desafortunadamente, estas oportunidades conllevan importantes desafíos éticos. En particular, el uso extensivo de cada vez más datos, a menudo personales (Big Data), la creciente dependencia de algoritmos para analizarlos con el fin de dar forma a las opciones y tomar decisiones (incluido el machine learning, la inteligencia artificial y la robótica), y la reducción gradual de la participación humana o la supervisión de procesos automáticos, plantean **cuestiones urgentes sobre la equidad, la responsabilidad y el respeto de los derechos humanos**.

Así, la **preferibilidad social** debe ser el principio rector para lograr un equilibrio ético sólido para cualquier proyecto digital con un impacto en la vida humana. Entonces, Luciano Floridi, filósofo de la Universidad de Oxford, entiende la ética digital como la rama de la ética que estudia y evalúa los problemas morales relacionados con:

- **la información y los datos**: incluida la generación, grabación, curación, procesamiento, difusión, intercambio y uso.
- **Los algoritmos**: incluida la inteligencia artificial, agentes artificiales, aprendizaje de máquinas y robots.
- **Y las prácticas e infraestructuras correspondientes** incluida la innovación responsable, la programación, la piratería, los códigos profesionales y los estándares.

Todo lo anterior, con el fin de formular y respaldar soluciones moralmente buenas. Por lo tanto, una forma exitosa de identificar y evaluar proyectos socialmente preferibles es analizarlos sobre la base de sus resultados. Estos, son exitosos en la medida en que ayudan a reducir, mitigar o erradicar un determinado problema social o ambiental, sin introducir nuevos daños ni amplificar los existentes. Así, Cowls (2021) continúa sugiriendo que:

un proyecto basado en IA que es socialmente preferible, se define formalmente como: aquel que es diseñado, desarrollado y desplegado de manera que ayude a **1.** Prevenir, mitigar y / o resolver problemas que afecten negativamente la vida humana y / o el bienestar del mundo natural, y / o **2.** Permite desarrollar actividades socialmente deseables o ambientalmente sostenibles, y a su vez **3.** No introduce nuevas formas de daño y / o amplifica las inequidades existentes.

3.1 Ética de la Inteligencia Artificial

Los algoritmos se han convertido en agentes cruciales para toda la infraestructura digital en la que actualmente habitamos las sociedades digitales y de la información. Desde sistemas de recomendación, filtrado de información, optimización de servicios y procesos, hasta interpretación de variables fisiológicas en entornos deportivos de alto rendimiento y médicos de alta complejidad. Al día de hoy, gobiernos, instituciones educativas, empresas, cortes de justicia, hospitales, dependen de sistemas de IA para tomar decisiones cada vez más cruciales. Y si bien, es claro que la delegación de tareas a los sistemas de IA puede mejorar la eficiencia y permitir nuevas soluciones; también lo es que, estos beneficios van acompañados de desafíos éticos (Mökeler et. al, 2021). Por ejemplo, estas tecnologias pueden producir resultados discriminatorios, violar la privacidad individual y socavar la autodeterminación humana (Veliz, 2021; Zuboff, 2019).

Por lo tanto, se necesitan nuevos mecanismos de gobernanza y previsión (Floridi & Strait, 2020), que ayuden a las organizaciones a diseñar e implementar estas tecnologías de manera ética, al tiempo que permiten a la sociedad cosechar todos los beneficios económicos y sociales de la inteligencia artificial (Floridi & Cowls, 2019; Floridi, 2016**)**. Como resultado, muchas organizaciones han lanzado una amplia gama de iniciativas para establecer *principios éticos* para la adopción de IA socialmente preferible. Entre los que se encuentran: Los principios de Asilomar para la Inteligencia Artificial (2017), La Declaración para la Inteligencia Artificial Responsable (Montreal Declaration, 2017), Los principios ofrecidos por la IEEE en su artículo Diseño Éticamente Alineado (IEEE, 2017), Los principios éticos ofrecidos por el Grupo de Europeo de Ética en Robótica (EGE, 2018) y los 5 Principios Éticos Generales para un código de la Inteligencia Artificial de House of Lords de Reino Unido (House of Lords, 2018). Afortunadamente, entre los 5 conjuntos de principios hay una

buena superposición, lo cual evidencia que los valores democráticos están presentes, en donde los derechos humanos sirven como guía (Floridi et. al., 2019)**.**

Y cuando estos conjuntos se revisan a la luz de los 4 principios bioéticos: Beneficencia, No maleficencia, Autonomía y Justicia, se ve con mayor claridad la sintonía que hay entre los conjuntos de principios (Floridi, 2020). Pero de acuerdo al trabajo comparativo realizado por el Digital Ethics Lab en cabeza de Luciano Foridi, se encuentra que es necesario añadir un nuevo principio que vele por la *inteligibilidad* y la *responsabilidad* de las acciones llevadas a cabo por los sistemas de IA. Debido a esto, Floridi y Cowls (2019) sugieren la *explicabilidad* como el nuevo principio que puede complementar a la bioética para hacerle frente a los retos traídos por la revolución digital.

Así, a la luz de estos hallazgos, las siguientes son las descripciones no exhaustivas de los principios para la IA:

- **Beneficencia:** es promover el bienestar, preservar la dignidad y cuidar el planeta con los potenciales que ofrece la IA.
- **No maleficencia:** aunque "hacer el bien" (beneficencia) y "no hacer daño" (no maleficencia) pueden parecer lógicamente equivalentes, no lo son y representan principios distintos. Es de especial preocupación la prevención de infracciones en privacidad, seguridad, y la capacidad de ser cautos.
- **Autonomía:** como se ha mencionado anteriormente, al implementar la IA, cedemos voluntariamente parte de nuestro poder de toma de decisiones a estos artefactos tecnológicos. Por lo tanto, sugerir el principio de autonomía en el contexto de la IA significa lograr un equilibrio entre el poder de toma de decisiones que mantenemos para nosotros y el que delegamos a los agentes artificiales.
- **Justicia:** La IA deberá contribuir a la justicia global y al acceso equitativo a los beneficios de las tecnologías digitales. Promoviendo la prosperidad y preservando la solidaridad.
- **Explicabilidad:** la adición del principio de explicabilidad está sustentada en 2 objetivos: 1. en **la inteligibilidad**, que busca responder la pregunta por *¿cómo funciona el sistema?*. Y, 2. **la responsabilidad**, en el sentido ético, pues busca elaborar sobre la pregunta *¿quién responde por la forma en que funciona el sistema?*

Entonces, al disponer de los anteriormente mencionados principios éticos, ¿cómo se pueden implementar en el proceso de desarrollo y diseño de software, algoritmos y sistemas de IA? Además, ¿cómo se puede confirmar y verificar que realmente se materialicen estos principios en los sistemas de IA? Estas dos preguntas servirán de guía para las próximas 2 secciones, pues la primera hace referencia a los métodos y medios éticos que permiten el diseño y el despliegue de sistemas de IA socialmente preferibles. Y la segunda, a los mecanismos de gobernanza que procurarán garantizar que los sistemas de IA que se desplieguen en la sociedad han sido diseñados con unos mínimos éticos y que demuestran una continua revisión y actualización a la luz de sus propósitos y efectos.

3.1.1 Ethics as a service: orientación ética para el diseño de Inteligencia Artificial socialmente preferible

El concepto *ética como servicio (Ethics as a service)* propone la disposición de mediación por parte de las directivas y desarrolladores de software de una organización, para que a la hora de diseñar sistemas de IA construyan en torno a tecnologías y soluciones socialmente preferibles. En específico, hay dos tensiones que se deben evaluar constantemente en la búsqueda de sistemas de IA éticamente sólidos. La tensión entre **demasiado flexible y demasiado estricto**, apunta a la mediación que las circunstancias en términos normativos, de principios, reglas generales para instancias específicas. Pues como Ananny & Crawford (2018) lo señalan, los sistemas algorítmicos son ensamblajes entre agentes humanos y no-humanos que tienen muchos impactos no determinísticos, es por esto que se requiere desarrollar el discernimiento para poder llegar a una comprensión de cómo funciona toda la infraestructura donde opera el sistema de IA y su relación con las leyes, regulaciones y normas. La otra tensión surge entre **la responsabilidad delegada y la responsabilidad centralizada,** que hace referencia a cuándo solicitar auditoría externa y qué aspectos auditar con el auditor interno. Recordando que la auditoría ética externa debe ser un componente central para cualquier forma de operacionalización de la ética de la IA, siendo este un proceso colaborativo de continua negociación. Además, es importante resaltar que las auditorías internas, siendo de gran valor, pueden presentar conflictos de interés que conlleven a la incapacidad de mantener una opinión objetiva por parte del auditor; mientras que las auditorías externas cuentan con limitaciones como no tener acceso a toda la información relevante sobre el sistema por términos contractuales o de privacidad.

Para términos ilustrativos, se presentará el marco ético propuesto por Digital Catapult (2020) y que más que un marco es un dispositivo metodológico que busca la mediación a la hora de implementar los principios éticos en los sistemas de IA, además de revisar otros aspectos más prácticos y de igual relevancia que inciden en cómo funcionará el sistema de IA en la sociedad. El marco fue creado por el Comité de Ética de Digital Catapult, y consta de siete conceptos, cada uno con las preguntas correspondientes que tienen como objetivo informar cómo estos conceptos podrían aplicarse en la práctica. Este dispositivo funciona haciendo énfasis en las preguntas más que en los principios éticos, porque las preguntas ayudan a esclarecer dónde se deben considerar los principios en la práctica, y las preguntas no asumen una respuesta "correcta" universal. Morley (2020) describe así el dispositivo:

> Este consta de cuatro niveles. El *primer nivel*, consiste en los cinco principios de alto nivel identificados mencionados por Floridi et al., (2018): **beneficencia, no maleficencia, autonomía, justicia, explicabilidad.** El *segundo nivel* consta de siete interpretaciones (o definiciones contextuales) de estos principios identificados a través de consultas de análisis documental, con profesionales en IA externos y de la organización, además de otros stakeholders, afectados por el sistema de IA. El *tercer nivel* operacionaliza el concepto de ética del discurso de Habermas (Buhmann et al., 2019), es decir, un enfoque que busca establecer valores normativos y verdades éticas a través del discurso abierto y la dialéctica, que consta de una serie de preguntas que están diseñadas para motivar a los profesionales de la IA a realizar análisis de prospectiva ética (Floridi & Strait, 2020). El *cuarto nivel* proporciona acceso a herramientas más prácticas y menos discursivas, p. Ej. Bibliotecas de Python diseñadas para identificar sesgos en los datos.

Los siguientes son los 7 conceptos sobre los que se hacen exploraciones en el *segundo nivel*, para a través de la ética discursiva de Habermas, relacionados al sistema de IA y sus implicaciones sociales:

1. **Beneficios claros:** busca lograr claridad sobre los beneficios del producto o servicio

2. **Conocer y gestionar los riesgos:** se debe considerar la seguridad y el daño potencial, tanto como consecuencia del uso previsto del producto como de otros usos razonablemente previsibles.

3. **Utilizar los datos de forma responsable:** el cumplimiento de la legislación (como el General Data Protection Regulation) es un buen punto de partida para una evaluación ética de los datos y la privacidad. Aunque hay que tener en cuenta otros temas de relevancia como, la proveniencia de los datos, la aptitud de los datos y sus posibles sesgos representacionales.

4. **Ser merecedor de confianza:** las empresas deben poder explicar el propósito y las limitaciones de sus sistemas de IA para que los usuarios no sean engañados o confundidos.

5. **Diversidad, igualdad e inclusión:** las empresas deben considerar el impacto y la utilidad de sus sistemas de IA para las personas, los grupos más grandes y la sociedad en su conjunto, incluido el impacto en la ampliación o reducción de la desigualdad, permitiendo o restringiendo la discriminación y otros factores políticos, culturales y ambientales.

6. **Comunicación transparente:** las empresas deben poder comunicar claramente los beneficios y riesgos potenciales de sus sistemas de IA y las acciones que han tomado para generar beneficios y evitar, minimizar o mitigar los riesgos.

7. **Modelo de negocio:** las empresas deben considerar qué estructuras y procesos están empleando para generar ingresos u otro valor material para la organización a partir de sus sitemas de IA, ya que ciertos modelos comerciales o estrategias de precios pueden resultar en discriminación.

Entonces, en las sociedades digitales y de la información más avanzadas en donde la ética de la IA sirva como infraestructura para el apropiado diseño de sistemas de IA socialmente preferible, debe disponer de una metodología práctica que permita la definición e implementación de los principios éticos en el desarrollo de los sistemas, y además, de un sistema de auditoría que permita la revisión iterativa y constante de la evolución de los sistemas.

3.1.2 La auditoría basada en la ética como mecanismo de gobernanza de la Inteligencia Artificial

Además de los códigos y marcos éticos mencionados anteriormente, la proliferación de este tipo de documentos está en aumento, poniendo en evidencia que existe una brecha significativa entre la teoría de los principios éticos de la IA y la aplicación práctica de estos en el diseño de los sistemas de IA (Morley et. al., 2021). Es por esto que se hace necesario encontrar maneras para implementar de forma efectiva los principios definidos por los códigos y marcos éticos, y a su vez, mecanismos de revisión y gobernanza constante que permitan verificar la apropiada materialización de los principios en los sistemas de IA. Así, siguiendo con la línea de investigación desarrollada por el Oxford Internet Institute, se propone la *auditoría basada en la ética* como el mecanismo de gobernanza que apunta a cerrar la brecha entre el *qué* (principios éticos) y el *cómo* (implementación) de la ética de la IA, sirviendo como mecanismo de gobernanza para la verificación de la implementación de los principios éticos.

Es importante resaltar que la auditoría basada en la ética no busca mecanizar la ética; lo que principalmente busca es ayudar a identificar, visualizar y comunicar los valores éticos y morales que están enmarcados en un sistema de IA (Morley et al., 2019). Y para lograr este objetivo, procura abrir la conversación entre desarrolladores de software y administradores/gerentes de las organizaciones que diseñan los sistemas. Es por esto que hay que enfatizar que la responsabilidad principal de identificar y ejecutar los pasos para garantizar que los sistemas de IA sean éticamente sólidos recae en la administración de las organizaciones que diseñan y despliegan estos sistemas (Cowls J., 2020).

Así pues, la auditoría basada en la ética, se concibe como un mecanismo dialéctico, donde el auditor se encarga de procurar que las preguntas correctas se aborden en el proceso de diseño de los sistemas de IA, y que sean respondidas adecuadamente (Morley et al., 2021). Pero además, en el contexto propiciado por el ejercicio de auditoría, la conversación comienza desde los principios (abstractos) y se dirige hacia la intervención directiva/gerencial (implementación) a través del ciclo del producto, lo cual busca permear la conceptualización, el diseño, el despliegue y el uso de los sistemas de IA. Adicionalmente, debido a la naturaleza cambiante de los sistemas de IA, el ejercicio de auditoría debe ser continuo, puesto que más que buscar sistemas sin sesgo alguno o perfectos; lo que busca es tener claridad respecto a cuando el sistema puede estar causando daño o se está comportando

de formas inesperadas (Cowls J, 2020). Esto se hace a través de programas de supervisión y documentación de las características de desempeño de una manera comprensible (Mitchell et al., 2019).

Para Mökander y colaboradores (2021), el ejercicio de auditoría basada en la ética coordinado por el auditor externo, se encarga de guiar conversaciones en 3 grandes fases del ciclo de vida del sistema de IA y que se repiten de forma iterativa, a saber:

1. **Auditoría de funcionalidad:** en donde se comprende y abre la conversación respecto a la organización que desarrolla el sistema de IA y el diseño conceptual del sistema.
    - **Sobre la organización**, se tiene en cuenta la visión, el código de ética y la experiencia y formación de los empleados y diseñadores.
    - **Sobre la conceptualización del sistema,** se tiene en cuenta la definición del uso que se le dará al sistema, las especificaciones del sistema, las verificaciones de cumplimiento, y la revisión de las conversaciones con los grupos sociales que interactúan o se verán afectados por el sistema.

2. **Auditoría del código:** en donde se abordarán el desarrollo y la evaluación técnica del código.
    - **Sobre el desarrollo:** se analizará la alineación ética del diseño del sistema, las fuentes de los datos, el entrenamiento del modelo, y la estructura de actualización y rediseño del modelo.
    - **Sobre la evaluación técnica:** se abordarán el testeo, la definición y manejo de los riesgos, la ficha técnica y de verificación, y el despliegue del sistema.

3. **Auditoría del impacto:** en donde se revisarán y discutirán, de nuevo la evaluación técnica y la operación del sistema:
    - **Operación:** se analizan y monitorizan los resultados y efectos del sistema, las revisiones, se dará manejo a las quejas y reclamos que ha causado el sistema, se planificaran las próximas auditorías, y los planes de mantenimiento.

Este es pues, un proceso iterativo y no un destino, que procura informar, organizar e interconectar las estructuras de gobernanza de la organización implicadas.

4 Conclusiones

Nos encontramos pues, en sociedades que cada vez dependen más de las tecnologías digitales para su funcionamiento. Entre las nuevas tecnologías encontramos los sistemas de IA mencionados anteriormente y las fuentes de datos necesarias para su funcionamiento, brindando un poder de gran magnitud a sus diseñadores, que puede afectar desde individuos hasta sistemas sociales-ambientales enteros.

Es por esto que dispositivos organizacionales como la *ética como servicio* son cada vez más necesarios para la apropiada materialización de los principios éticos en el diseño e implementación de sistemas de IA. Pero además, sistemas de gobernanza y control como la *auditoría basada en la ética* se convierten en pilares de la infraestructura ética (Floridi, 2018) en las sociedades digitales, con el objetivo de diseñar sistemas y soluciones **socialmente preferibles.** Es claro que este es un campo que se encuentra en las primeras etapas de su desarrollo y que todavía necesita de tiempo y experimentación para identificar las limitaciones que dispositivos y prácticas como *la ética como servicio* y *la auditoría basada en la ética,* puedan tener.

Aun así, estos son dispositivos pioneros que continúan construyendo hacia la gobernanza de lo digital, fundamentando las decisiones de los sistemas de IA al brindar herramientas de visualización y monitoreo a los resultados de estos, procurando trascender el modelo black box. También, una vez implementados interna y externamente por organizaciones y gobiernos, permitirán mantener informados a los ciudadanos y usuarios de cómo las decisiones a las que un sistema de IA ha llegado, permitiendo someterlas a una nueva consideración. Buscando aliviar el sufrimiento humano y anticipar el posible daño causado por estas tecnologías; pero además, y más importante aún, asignando las responsabilidades a las personas y organizaciones adecuadas.

5 Referencias

- Bird, Sarah and Barocas, Solon and Crawford, Kate and Diaz, Fernando and Wallach, Hanna, Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI (October 2, 2016). Workshop on Fairness, Accountability, and Transparency in Machine Learning, 2016, Available at SSRN: https://ssrn.com/abstract=2846909
- Burr C., Morley J. (2020) Empowerment or Engagement? Digital Health Technologies for Mental Healthcare. In: Burr C., Milano S. (eds) The 2019 Yearbook of the Digital Ethics Lab. Digital Ethics Lab Yearbook. Springer, Cham. https://doi-org.ezproxy.eafit.edu.co/10.1007/978-3-030-29145-7_5
- Burr, C., Cristianini, N. & Ladyman, J. An Analysis of the Interaction Between Intelligent Software Agents and Human Users. Minds & Machines 28, 735–774 (2018). https://doi.org/10.1007/s11023-018-9479-0
- Burr, C., Cristianini, N. Can Machines Read our Minds?. Minds & Machines 29, 461–494 (2019). https://doi.org/10.1007/s11023-019-09497-4
- Cath C. (2019) Internet Governance and Human Rights: A Literature Review. In: Öhman C., Watson D. (eds) The 2018 Yearbook of the Digital Ethics Lab. Digital Ethics Lab Yearbook. Springer, Cham. https://doi.org/10.1007/978-3-030-17152-0_8
- Crawford, Kate and Lumby, Catharine, Networks of Governance: Users, Platforms, and the Challenges of Networked Media Regulation (February 6, 2013). International Journal of Technology Policy and Law, Vol. 2 No. 1 (2013, Forthcoming), Available at SSRN: https://ssrn.com/abstract=2246772
- Cowls J. (2020) Deciding How to Decide: Six Key Questions for Reducing AI's Democratic Deficit. In: Burr C., Milano S. (eds) The 2019 Yearbook of the Digital Ethics Lab. Digital Ethics Lab Yearbook. Springer, Cham. https://doi-org.ezproxy.eafit.edu.co/10.1007/978-3-030-29145-7_7
- Cowls, J., Tsamados, A., Taddeo, M. et al. A definition, benchmark and database of AI for social good initiatives. Nat Mach Intell 3, 111–115 (2021). https://doi.org/10.1038/s42256-021-00296-0
- De Cremer, D., Kasparov, G. The ethical AI—paradox: why better technology needs more and not less human responsibility. AI Ethics (2021). https://doi.org/10.1007/s43681-021-00075-y

- Floridi L. (2019) The Green and the Blue: Naïve Ideas to Improve Politics in a Mature Information Society. In: Öhman C., Watson D. (eds) The 2018 Yearbook of the Digital Ethics Lab. Digital Ethics Lab Yearbook. Springer, Cham. https://doi.org/10.1007/978-3-030-17152-0_12
- Floridi L. (2020) What the Near Future of Artificial Intelligence Could Be. In: Burr C., Milano S. (eds) The 2019 Yearbook of the Digital Ethics Lab. Digital Ethics Lab Yearbook. Springer, Cham. https://doi-org.ezproxy.eafit.edu.co/10.1007/978-3-030-29145-7_9
- Floridi L., Cath C., Taddeo M. (2019) Digital Ethics: Its Nature and Scope. In: Öhman C., Watson D. (eds) The 2018 Yearbook of the Digital Ethics Lab. Digital Ethics Lab Yearbook. Springer, Cham. https://doi.org/10.1007/978-3-030-17152-0_2
- Floridi, L. (2007). A Look into the Future Impact of ICT on our Lives. SSRN Electronic Journal.
- Floridi, L. (2012). Hyperhistory and the Philosophy of Information Policies. Philosophy & Technology, 25(2), 129–131.
- Floridi, L. (2015). Ethics in the Age of Information. Lecture for The Alan Turing Institute.
- Floridi, L. (2021). Luciano Floridi and the technological gambit. https://www.eni.com/en-IT/global-energy-scenarios/luciano-floridi-technological-gambit.html.
- Floridi, L. Children of the Fourth Revolution. Philos. Technol. 24, 227 (2011). https://doi.org/10.1007/s13347-011-0042-7
- Floridi, L. Digital's Cleaving Power and Its Consequences. Philos. Technol. 30, 123–129 (2017). https://doi.org/10.1007/s13347-017-0259-1
- Floridi, L. Establishing the Rules for Building Trustworthy AI (May 07, 2019). Available at SSRN: https://ssrn.com/abstract=3858392 or http://dx.doi.org/10.2139/ssrn.3858392
- Floridi, L. Hyperhistory and the Philosophy of Information Policies (May 8, 2012). Philosophy & Technology volume 25, pages 129–131 (2012), Available at SSRN: https://ssrn.com/abstract=3854425
- Floridi, L. Open Data, Data Protection, and Group Privacy. Philos. Technol. 27, 1–3 (2014). https://doi.org/10.1007/s13347-014-0157-8
- Floridi, L. Robots, Jobs, Taxes and Responsibilities (March 14, 2017). Available at SSRN: https://ssrn.com/abstract=3843510 or http://dx.doi.org/10.2139/ssrn.3843510

- Floridi, L. Technological Unemployment, Leisure Occupation, and the Human Project (May 06, 2014). Available at SSRN: https://ssrn.com/abstract=3843543 or http://dx.doi.org/10.2139/ssrn.3843543
- Floridi, L. The European Legislation on AI: A Brief Analysis of its Philosophical Approach (June 1, 2021). Available at SSRN: https://ssrn.com/abstract=3873273 or http://dx.doi.org/10.2139/ssrn.3873273
- Floridi, L. The Unsustainable Fragility of the Digital, and What to Do About It. Philos. Technol. 30, 259–261 (2017). https://doi.org/10.1007/s13347-017-0280-4
- Floridi, L., (2021, May 24). La clave es gobernar lo digital. EL PAÍS. https://elpais-com.cdn.ampproject.org/c/s/elpais.com/opinion/2021-05-24/la-clave-es-gobernar-lo-digital.html?outputType=amp.
- Floridi, L., Cowls, J., King, T.C. et al. How to Design AI for Social Good: Seven Essential Factors. Sci Eng Ethics 26, 1771–1796 (2020). https://doi.org/10.1007/s11948-020-00213-5
- Floridi, L., Strait, A. Ethical Foresight Analysis: What it is and Why it is Needed?. Minds & Machines 30, 77–97 (2020). https://doi.org/10.1007/s11023-020-09521-y
- Floridi, Luciano and Cowls, Josh, A Unified Framework of Five Principles for AI in Society (September 20, 2019). Available at SSRN: https://ssrn.com/abstract=3831321 or http://dx.doi.org/10.2139/ssrn.3831321
- Frischen, K. (2021, April 26). 'Make Algorithmic Audits As Ubiquitous As Seatbelts'-Why Tech Needs Outside Help To Serve Humanity. Forbes. https://www.forbes.com/sites/ashoka/2021/04/26/make-algorithmic-audits-as-ubiquitous-as-seatbeltswhy-tech-needs-outside-help-to-serve-humanity/?sh=179935a149a1.
- Hao, K. (2020, December 8). The coming war on the hidden algorithms that trap people in poverty. MIT Technology Review. https://www.technologyreview.com/2020/12/04/1013068/algorithms-create-a-poverty-trap-lawyers-fight-back/?utm_medium=tr_social&utm_campaign=site_visitor.unpaid.engagement&utm_source=Twitter#Echobox=1613602785.
- Hao, K. (2021, June 15). Inside the fight to reclaim AI from Big Tech's control. MIT Technology Review. https://www.technologyreview.com/2021/06/14/1026148/ai-big-tech-timnit-gebru-paper-ethics/.
- Jake Metcalf and Kate Crawford, 2016 'Where are the Human Subjects in Big Data Research? The Emerging Ethics Divide,' Big Data & Society, special issue on Critical

Data Studies, Spring 2016. http://bds.sagepub.com/content/3/1/2053951716650211.full.pdf+htm

- Jon Whittle. (2021, March 31). AI can now learn to manipulate human behaviour. The Conversation. https://theconversation.com/ai-can-now-learn-to-manipulate-human-behaviour-155031?fbclid=IwAR0YDYGl1Jb-ctw39fAuL5nQS7fSSoUdyRH9APvyOimSdMyCjeqhVhSeSDI.
- Joseph B. Bak-Coleman, Mark Alfano, Wolfram Barfuss, Carl T. Bergstrom, Miguel A. Centeno, Iain D. Couzin, Jonathan F. Donges, Mirta Galesic, Andrew S. Gersick, Jennifer Jacquet, Albert B. Kao, Rachel E. Moran, Pawel Romanczuk, Daniel I. Rubenstein, Kaia J. Tombak, Jay J. Van Bavel, Elke U. Weber.
- Kate Crawford and Jason Schultz, 2019 'AI Systems As State Actors', Columbia Law Review, 119(7), 1941-1972. https://columbialawreview.org/content/ai-systems-as-state-actors/
- Kate Crawford and Vladan Joler, 2018 'Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources', AI Now Institute and Share Lab, September 7, 2018. https://anatomyof.ai
- Lauer, D. Facebook's ethical failures are not accidental; they are part of the business model. AI Ethics (2021). https://doi.org/10.1007/s43681-021-00068-x
- Milligan, M. (2018). Technology and the Ethics GapApr 12, 2018. ABET. https://www.abet.org/technology-and-the-ethics-gap/.
- Mittelstadt B, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: Mapping the debate. Big Data & Society. December 2016. doi:10.1177/2053951716679679
- Mittelstadt, Brent, Principles Alone Cannot Guarantee Ethical AI (May 20, 2019). Nature Machine Intelligence, November 2019, Available at SSRN: https://ssrn.com/abstract=3391293 or http://dx.doi.org/10.2139/ssrn.3391293
- Mökander, J., Floridi, L. Ethics-Based Auditing to Develop Trustworthy AI. Minds & Machines 31, 323–327 (2021). https://doi.org/10.1007/s11023-021-09557-8
- Mökander, J., Morley, J., Taddeo, M. et al. Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. Sci Eng Ethics 27, 44 (2021). https://doi.org/10.1007/s11948-021-00319-4
- Mökander, J., Schroeder, R. AI and social theory. AI & Soc (2021). https://doi.org/10.1007/s00146-021-01222-z

- Morley, J., Elhalal, A., Garcia, F. et al. Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. Minds & Machines 31, 239–256 (2021). https://doi.org/10.1007/s11023-021-09563-w
- Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3830348
- Morley, J., Morton, C., Karpathakis, K., Taddeo, M., & Floridi, L. (2021). Towards a Framework for Evaluating the Safety, Acceptability and Efficacy of Ai Systems for Health: An Initial Synthesis. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3826358
- Richardson, Rashida and Schultz, Jason and Crawford, Kate, Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice (February 13, 2019). 94 N.Y.U. L. REV. ONLINE 192 (2019), Available at SSRN: https://ssrn.com/abstract=3333423
- Öhman C., Watson D. (2019) Digital Ethics: Goals and Approach. In: Öhman C., Watson D. (eds) The 2018 Yearbook of the Digital Ethics Lab. Digital Ethics Lab Yearbook. Springer, Cham. https://doi.org/10.1007/978-3-030-17152-0_1
- Proceedings of the National Academy of Sciences Jul 2021, 118 (27) e2025764118; DOI: 10.1073/pnas.2025764118
- Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51-59.
- Sunstein, C. (2008). Democracy and the Internet. In J. Van den Hoven & J. Weckert (Eds.), Information Technology and Moral Philosophy (Cambridge Studies in Philosophy and Public Policy, pp. 93-110). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511498725.006
- Taddeo, M. Philosophy and Computing in Information Societies. Minds & Machines 26, 203–204 (2016). https://doi.org/10.1007/s11023-016-9400-7
- Taddeo, M. The Civic Role of Online Service Providers. Minds & Machines 29, 1–7 (2019). https://doi.org/10.1007/s11023-019-09495-6
- Taddeo, M. Trust in Technology: A Distinctive and a Problematic Relation. Know Techn Pol 23, 283–286 (2010). https://doi.org/10.1007/s12130-010-9113-9
- Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. Science, 361(6404), 751–752. https://doi.org/10.1126/science.aat5991

- Taddeo, M., McCutcheon, T. & Floridi, L. Trusting artificial intelligence in cybersecurity is a double-edged sword. Nat Mach Intell 1, 557–560 (2019). https://doi.org/10.1038/s42256-019-0109-1
- Taddeo, M., Tsamados, A., Cowls, J., & Floridi, L. (2021). Artificial Intelligence and the Climate Emergency: Opportunities, Challenges, and Recommendations. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3873881
- The Onlife Initiative (2015) The Onlife Manifesto. In: Floridi L. (eds) The Onlife Manifesto. Springer, Cham. https://doi.org/10.1007/978-3-319-04093-6_2
- Tsamados, Andreas and Aggarwal, Nikita and Cowls, Josh and Morley, Jessica and Roberts, Huw and Taddeo, Mariarosaria and Floridi, Luciano, The Ethics of Algorithms: Key Problems and Solutions (July 28, 2020). Available at SSRN: https://ssrn.com/abstract=3662302 or http://dx.doi.org/10.2139/ssrn.3662302
- Turilli, Matteo (2008). Ethics and the Practice of Software Design. In P. Brey, A. Briggle & K. Waelbers (eds.), _Current Issues in Computing and Philosophy_. IOS Press.
- Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C. & Tucker, J. A., Political psychology in the digital (mis)Information age: A model of news belief and sharing, Journal of Psychological Study of Social Issues, 15, 84– 113.
- Veliz, C. (2021). Privacy is Power. Random House UK.
- Venkataramakrishnan, S. (2021, January 26). Human rights come under pressure from digital controls. Financial Times. https://www.ft.com/content/b7db783d-bf3c-4940-87ca-5a057dd71464.
- Vogl, T.M., Seidelin, C., Ganesh, B. and Bright, J. (2020), Smart Technology and the Emergence of Algorithmic Bureaucracy: Artificial Intelligence in UK Local Authorities. Public Admin Rev, 80: 946-961. https://doi.org/10.1111/puar.13286
- Wachter, Sandra and Mittelstadt, Brent and Floridi, Luciano, Transparent, Explainable, and Accountable AI for Robotics (May 31, 2017). Science Robotics, Vol. 2, Issue 6, eaan6080, 31 May 2017, DOI:10.1126/scirobotics.aan6080, Available at SSRN: https://ssrn.com/abstract=3011890
- Wachter, Sandra and Mittelstadt, Brent and Russell, Chris, Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI (March 3, 2020). Computer Law & Security Review (forthcoming), Available at SSRN: https://ssrn.com/abstract=3547922 or http://dx.doi.org/10.2139/ssrn.3547922

- Wachter, Sandra, The GDPR and the Internet of Things: A Three-Step Transparency Model (February 5, 2018). Law, Innovation and Technology doi.org/10.1080/17579961.2018.1527479, Available at SSRN: https://ssrn.com/abstract=3130392 or http://dx.doi.org/10.2139/ssrn.3130392
- Watson D. (2020) The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. In: Burr C., Milano S. (eds) The 2019 Yearbook of the Digital Ethics Lab. Digital Ethics Lab Yearbook. Springer, Cham. https://doi-org.ezproxy.eafit.edu.co/10.1007/978-3-030-29145-7_4
- Watson, D. (2021, April 8). AI Researcher: Stop Calling Everything "Artificial Intelligence". Mind Matters. https://mindmatters.ai/2021/04/ai-researcher-stop-calling-everything-artificial-intelligence/.
- Why We Should End the Data Economy. The Reboot. (2021, June 10). https://thereboot.com/why-we-should-end-the-data-economy/.
- Zhang, J. C. and D., Jensen, B., Lynch, S., & Waikar, S. (2021, March 3). AI Index 2021. Stanford HAI. https://hai.stanford.edu/research/ai-index-2021?utm_source=twitter&utm_medium=social&utm_content=Stanford+HAI_twitter_StanfordHAI_202103030815_sf139686683&utm_campaign=&sf139686683=1.
- Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, et al. (2017) Ten simple rules for responsible big data research. PLoS Comput Biol 13(3): e1005399. https://doi.org/10.1371/journal.pcbi.1005399
- Zuboff, S. (2020). The age of surveillance capitalism: the fight for a human future at the new frontier of power. Public Affairs.