Documento de trabajo 1/2020
9 de enero de 2020

REAL INSTITUTO elcano ROYAL INSTITUTE

Geopolítica de la ética en Inteligencia Artificial

Andrés Ortega Klein

Geopolítica de la ética en Inteligencia Artificial

Andrés Ortega Klein | Investigador senior asociado, Real Instituto Elcano | @andresortegak

Índice

Resumen

Los principios que han de regir la ética en la Inteligencia Artificial (IA) son objeto de una multiplicidad de propuestas por parte de Estados, organizaciones internacionales y profesionales, y grandes empresas. Detrás hay una carrera geopolítica, con la idea de que estos principios –sobre los que hay un acuerdo bastante amplio, como se ha plasmado en el G20– se han de incorporar, en lo que es la parte más difícil, en una programación que siempre ha de quedar bajo control del ser humano. Detrás de las principales propuestas que resumimos y analizamos, hay conflictos de cultura, valores y control, que cobrarán más importancia a medida que se acerque la fase de la reglamentación práctica.

(1) Introducción

Numerosos gobiernos, organizaciones internacionales y profesionales, además de empresas, han y siguen presentando propuestas sobre ética en la Inteligencia Artificial (IA). A principios de 2019 ya había casi 90 de ellas. Algunas, como la de la OCDE (Organización para la Cooperación y Desarrollo Económicos), la del G20, basada en la anterior, o las del IEEE (Instituto de Ingeniería Eléctrica y Electrónica) –quizá la más

completa–,[1] agrupan a muy diversos países o profesionales de distintas culturas. Incluso China, en lo que parece un intento de acercarse al debate occidental, ha producido sus propuestas de ética para la IA,[2] y ha suscrito la del G20.[3] Pues detrás de esta nube de propuestas hay también una cuestión de influencia y de poder, de geopolítica, que irá a más cuando se pase de los principios a su aplicación.

Esta cuestión de la ética en la IA se engloba dentro de una temática más amplia sobre el control de la tecnología por los propios humanos. Por ejemplo, el pasado junio, el panel sobre cooperación digital del secretario general de la ONU, co-presidido por Melinda Gates y Jack Ma (entonces aún al frente de Alibaba), produjo un informe y una Declaración de Interdependencia Digital, que pedía a todos los *stakeholders* "colaborar en nuevas formas de lograr una visión del futuro de la humanidad en la que se utilicen tecnologías digitales asequibles y accesibles para permitir el crecimiento económico y las oportunidades sociales, disminuir la desigualdad, mejorar la paz y la seguridad, promover el medio ambiente la sostenibilidad, preservar el albedrío humano y promover los derechos humanos".[4]

La ética de la IA se ha convertido en un tema de debate global, no sólo entre expertos. No ha llegado aún al estadio de la regulación, pero sienta las bases para su futuro, y ese va a ser el momento de la verdad.

Hay una amplia coincidencia en las diversas propuestas. Pero tras la coincidencia en términos se pueden esconder divergencias en la práctica presente y futura. Hay también diferencias culturales importantes que conviene tener en cuenta en la búsqueda de unos patrones comunes y universales para la ética de la IA.

¿Qué se entiende por ética? Wikipedia la define como "disciplina filosófica que estudia el bien y el mal y sus relaciones con la moral y el comportamiento humano". En segundo lugar, como "conjunto de costumbres y normas que dirigen o valoran el comportamiento humano en una comunidad". El Diccionario de la Real Academia Española como "conjunto de normas morales que rigen la conducta de la persona en cualquier ámbito de la vida". Se podrían citar muchos autores. Nos limitaremos, como ejemplo, a Juan Manuel Orti y Lara, quien en su obra *Ética o principios de filosofia moral*, describía en 1853 la ética como "la ciencia que expone los principios de la moralidad de las acciones humanas y muestra cuáles son las que debe ejecutar el hombre en las diferentes relaciones en que se haya constituido".[5] En el caso que nos ocupa se trata de las

[1] Ethics in action, https://ethicsinaction.ieee.org/.

[2] Governance Principles for the New Generation Artificial Intelligence--Developing Responsible Artificial Intelligence (2019), http://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html?from=groupmessage&isappinstalled=0.

[3] G20 Ministerial Statement on Trade and Digital Economy, https://www.mofa.go.jp/files/000486596.pdf.

[4] Panel Launches Report & Recommendations For Building an Inclusive Digital Future (2019), https://digitalcooperation.org/panel-launches-report-recommendations/.

[5] Imprenta de las Escuelas Pías, Madrid.

personas que diseñan máquinas, las que las utilizan, probablemente también sus propietarios, y las propias máquinas y algoritmos que las conforman (¿robo-ética?).[6]

Conviene señalar que la ética, como la moral, no están conformadas por principios permanentes, sino que han ido cambiando a lo largo de diversas culturas y tiempos. Como también en la actualidad en lo referente a la IA.

A este estadio de la búsqueda y del debate, más allá de las buenas intenciones, se trata esencialmente de hacer nuestras intuiciones morales más explícitas y de construir una brújula moral para la IA. En un estadio en el que, como han indicado el emprendedor y psicólogo cognitivo Gary Marcus y el científico de computadores Ernest Davis, los sistemas de IA carecen de la capacidad para comprender conceptos de sentido común, como el tiempo, el espacio y la causalidad.[7] Aunque el objetivo es imbuir esta ética y este sentido común en los algoritmos –y muchos tecnólogos y científicos están trabajando en ello–, es aún una cuestión de suma dificultad técnica lograr que los algoritmos incorporen esos valores. También dependerá de los datos que se utilicen. El desarrollo de una ética de los datos será también interesante. ¿Ética para las máquinas o para las personas? ¿Ética para los programas o para los programadores? De momento, los esfuerzos van más dirigidos a las personas, a los que diseñan y son propietarios de esta IA, aunque con diversas tecnologías, la IA esté comenzando en algunos casos, en alguna algorítmica, a diseñarse a sí misma. Puede haber consecuencias no intencionadas en este debate sobre la ética en la IA.

Hay elementos novedosos respecto a las primeras propuestas de Isaac Asimov y sus leyes de la robótica. Las consideraciones éticas abarcan ahora cuestiones sociales (sobre todo la idea una tecnología inclusiva, al servicio de las personas, sin que nadie se quede atrás). E incluso a los posibles derechos de máquinas con IA “sensible”. O la especial inquietud, como puso de relieve el profesor Mori con su teoría “Del Valle Inquietante”,[8] que provocan humanoides demasiado realistas o incluso animaloides. Mori desaconsejaba hacer robots demasiado parecidos a los humanos. En todo caso, no es lo mismo darle una patata a un robot aspirador Roomba que a un perro Aibo de Sony: no hay la misma sensibilidad ante hechos parecidos. No se dan en todos los casos las mismas afinidades afectivas humano-máquina.

Dicho todo esto, hasta ahora, como indica Brent Mittelstadt,[9] del Oxford Internet Institute, las iniciativas han producido “declaraciones vagas, basadas en principios y valores de alto nivel, pero en la práctica pocas recomendaciones específicas y no logran abordar tensiones fundamentales normativas y políticas contenidas en conceptos clave (como equidad –*fairness*– o privacidad)”. Aunque algunas de las propuestas, como la

[6] La disciplina de la roboética ha sido introducida por Gianmarco Veruggio. Más sobre el tema en G. Veruggio y F. Operto (2008), “Roboethics: social and ethical implications of robotics”, en B. Siciliano y O. Khatib (eds.), *Springer Handbook of Robotics (1499-1524)*, Springer, Berlín/Heidelberg, https://link.springer.com/referenceworkentry/10.1007%2F978-3-540-30301-5_65#citeas.

[7] Gary Marcus y Ernest Davis (2019), “How to build Artificial Intelligence we can trust”, *The New York Times*, 6/IX9/2019, https://www.nytimes.com/2019/09/06/opinion/ai-explainability.html.

[8] Véase Andrés Ortega (2016), *La imparable marcha de los robots*, Alianza Editorial, Madrid.

[9] Brendt Mittelstadt (2019), “AI ethics – Too principled to fail?”, *Nature Machine Intelligence*, November 2019, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3391293.

iniciativa europea de la que hablaremos o la del IEEE, sí proporcionan una lista de preguntas que deberían ayudar a las empresas y sus empleados a abordar eventuales problemas ético-jurídicos de la IA.

Hay otro factor que conviene tener en cuenta en el actual debate: La cuestión de la ética es diferente para las grandes empresas que para las pequeñas. Las grandes está inmersas en este debate, y lo marcan. Las pequeñas no disponen de los medios suficientes, y seguirán, si acaso, las líneas generales que se vayan imponiendo.

Hay una cuestión de confianza en la ética de las empresas. Salvo alguna excepción, la realidad empresarial va aún por delante de la regulación por parte de las autoridades públicas. Según un informe de la consultora Capgemini,[10] el 86% de los directivos encuestados en varios países (como EEUU, el Reino Unido, Francia, Alemania y China) admiten haber observado prácticas éticamente cuestionables de IA en su empresa en los últimos dos o tres años. El informe considera que abordar las cuestiones éticas beneficiará a las organizaciones y su reputación. Es un factor positivo.

Pero de cara no sólo a la innovación en este terreno, sino a la confianza de los ciudadanos, hay una cuestión difícil de dilucidar: los dineros públicos y privados que se están poniendo en el tema y las organizaciones, institutos y programas que se están financiando con diversos fondos. Conviene distinguir entre los gestos "constructivos" (como el dinero de Wallenberg Foundation, el AI for Good Institute, los programas de investigación europeos Humane, etc.), los que aportan entidades públicas en países fiables o desde la propia UE y los que justifican cierta suspicacia porque parece que pretenden lavar la cara de los donantes más que realmente promover un uso ético de la IA –como los 350 millones de dólares de Blackstone a MIT para impulsar la computación por IA, los 27 millones de dólares de la Knight Foundation para el Fondo de Ética y Gobernanza de la Inteligencia Artificial, las iniciativas privadas de diversas empresas privadas o los fondos chinos para estos fines–.

Varias grandes empresas estadounidenses (como Microsoft[11] e IBM,[12] entre otras) están haciendo interesantes propuestas en materia de ética de la IA. También empresas europeas como Telefónica[13] han hecho propuestas constructivas en materia de ética y bienestar digital. Las empresas están muy presentes en los foros, en principio públicos de debate sobre la ética y la IA, lo que puede generar influencias y distorsiones, aunque también aportar exoperiencias desde la práctica.

Según *The New Statesman*,[14] Google ha gastado millones de euros sólo en su financiación a universidades británicas (17 millones de libras o 20 millones de euros sólo para la Universidad de Oxford). Google y DeepMind, perteneciente a la matriz de la

[10] Capgemini Research Institute (2019), "Why addressing ethical questions will benefit organizations", https://www.capgemini.com/es-es/wp-content/uploads/sites/16/2019/07/CRI-AI-in-Ethics_Web.pdf.
[11] Microsoft, "Microsoft AI Principles", https://www.microsoft.com/en-us/ai/our-approach-to-ai.

[12] IBM, "Everyday ethics for AI", https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.
[13] "Manifiesto por un Nuevo Pacto Digital", https://www.telefonica.com/manifiesto-digital/.

[14] Oscar Williams (2019), "How Big Tech funds the debate on AI ethics", *NewStateman*, 6/VI/2019, https://www.newstatesman.com/science-tech/technology/2019/06/how-big-tech-funds-debate-ai-ethics.

primera, Alphabet, han apoyado el trabajo en el Oxford Internet Institute (OII) sobre ética en IA, responsabilidad cívica de las empresas y otros aspectos. Esto no significa que no se deba hacer. Cuanto más *input*, mejor. Si se exige transparencia en la IA, también habría que exigirla para la financiación de los estudios sobre ética en IA.

También hay una cuestión de la enseñanza de la ética a los ingenieros, que era inexistente pero que se va abriendo paso en centros de enseñanza, entre otras razones debido a la demanda de los estudiantes. Centros como el MIT, Stanford y la Carnegie Mellon University ya cuentan con cursos específicos de ética, muy publicitados, también ante algunos escándalos que han surgido desde grandes empresas.[15] La ciencia ficción, pese a algunas exageraciones, puede ser una gran ayuda para estos estudiantes, pues en ella se adelantan muchos de los problemas éticos de la IA.

Está también la cuestión de los sesgos de varios tipos (genero, raza, cultura, edad, etc.). A menudo se construye la IA de acuerdo a lo que Seth Baum considera son visiones agregadas de la sociedad.[16]

Hay algunos problemas éticos que son específicos a la IA y otros más generales. Muchos de los que se suelen comentar en el debate público tienen que ver con la IA, pero porque amplifica problemas de fondo distintos como la mala estadística –sesgos–, ausencia de valores –micromarketing político– o la mediación por Internet –acceso a datos personales y agencia oculta (*hidden agency*)–. Lo que sí es la fuente de problemas éticos importantes específicos de la IA es la creciente autonomía de las entidades artificiales, y la solución más interesante es la de imbuir valores humanos en esas entidades artificiales. El tema lo afrontan bien el mencionado documento de la UE y del IEEE, que abordaremos más adelante.

Finalmente, está la cuestión crucial de cómo aplicar estos principios a la práctica de la IA, y el tema del *Value-Based Design* (VBD), diseño basado en valores. En general, la cuestión de la I+D "responsable" está muy relacionado con la ética y la IA.

(2) Antecedentes

En 1942 el escritor Isaac Asimov (1920-1992) acuñó sus famosas "tres leyes de la robótica":

(1) Un robot no hará daño a un ser humano ni, por inacción, permitirá que un ser humano sufra daño.

[15] Gregory Barber (2019), "What sci-fi can teach computer science about ethics", *Wired*, 26/VIII/2019, http://bit.ly/2knXgk9.
[16] Seth D. Baum, (2017), "Social choice ethics in artificial intelligence", *AI & Society*, https://doi.org/10.1007/s00146-017-0760-1. Ejemplos interesantes de algoritmos que discriminan pueden encontrarse en Jeffrey Dastin (2018), "Amazon scraps secret AI recruiting tool that showed bias against women", *Reuters*, 10/X/2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G y en Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner (2016), "Machine Bias", *ProPublica*, 23/V/2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

(2) Un robot debe obedecer las órdenes dadas por los seres humanos, excepto si estas órdenes entrasen en conflicto con la 1ª Ley.

(3) Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la 1ª o la 2ª Ley.

El autor de *Yo Robot* (1950) se percató de que podían entrar en contradicción o no funcionar, y añadió una cuarta.

(4) Un robot no puede hacer daño a la humanidad o, por inacción, permitir que la humanidad sufra daño.

Estas leyes robóticas son respuestas en teoría simples a problemas muy complejos. El propio Asimov exploró en sus escritos algunas de sus consecuencias inesperadas o no deseadas. No son morales, sino utilitarias. Y se basan en el supuesto de que los programadores pueden introducirlas en la IA y ésta no será capaz de superarlas de forma autónoma. Están pensadas por seres humanos para robots que entienden los propios humanos que los programan. Y para una IA limitada. No son fácilmente programables, entre otras cosas porque están escritas en lenguaje humano (originariamente en inglés). No en lenguaje máquina.

Pueden no bastar si se llega a la superinteligencia artificial (que superaría las mentes humanas más brillantes), concluyen en un trabajo el profesor Manuel Alfonseca, Manuel Cebrián y otros investigadores.[17] Pues, entonces, estos robots pueden llegar a movilizar ingentes recursos para fines que pueden resultar no sólo incomprensibles, sino incontrolables por los humanos, superando ampliamente a sus programadores. ¿Habría entonces que establecer otras reglas? ¿Quién las diseñará?

Stuart Russell, del Institute for the Future of Life, ha reformulado las leyes o reglas de Asimov para aplicarlas a lo que llama "IA probadamente beneficiosa" *(provably beneficial AI*):[18]

(a) El único objetivo del robot es maximizar la realización de valores humanos.
(b) El robot tiene una incertidumbre inicial sobre qué valores son esos.
(c) La mejor fuente de información sobre valores humanos es el comportamiento humano.

Como ejemplo más sofisticado y concreto, por su interés, citaremos también el de la Comisión de Ética sobre Conducción Automatizada –un tipo de robótica– impulsada por el Gobierno alemán, que llegó en 2017 a una serie de propuestas[19] que actualizan este

[17] Manuel Alfonseca, Manuel Cebrián, Antonio Fernandez-Anta, Lorenzo Coviello, Andres Abeliuk e Iyad Rahwany (2015), "Superintelligence cannot be contained", perspectiva presentada al *Journal of the Day Society Interface*, 12/X/2016.

[18] Stuart Russel (joint work with Dylan Hadfield-Menell, Smitha Milli, Anca Dragan, Pieter Abbeel, Tom Griffith), "Provably Beneficial AI", https://futureoflife.org/wp-content/uploads/2017/01/Stuart-Russell-conference.pdf?x90991.

[19] The Federal Government's action plan on the report by the Ethics Commission on Automated and (cont.)

problema y sus posibles soluciones, y que pueden inspirar futuras leyes para los robots. Sus elementos clave son:

- La conducción automatizada y conectada es un imperativo ético si los sistemas causan menos accidentes que los conductores humanos (balance de riesgo positivo).
- El daño a la propiedad debe tener prioridad sobre las lesiones personales. En situaciones peligrosas, la protección de la vida humana siempre debe tener la máxima prioridad.
- En el caso de situaciones de accidentes inevitables, cualquier distinción entre individuos basada en características personales (edad, género, constitución física o mental) es inadmisible.
- En cada situación de conducción, debe estar claramente regulado y ser evidente quién es responsable de la tarea de conducir: el humano o la computadora.
- Se debe documentar y almacenar quién conduce (para resolver posibles problemas de responsabilidad, entre otras cosas).
- Los conductores siempre deben poder decidir si los datos de sus vehículos se deben reenviar y usar (soberanía de los datos).

(3) Las principales propuestas

No pretendemos hacer aquí un repaso exhaustivo de todas las propuestas que se han hecho hasta la fecha o están en elaboración, sino seleccionar las de grupos de países o instituciones importantes o superpotencias de la IA como EEUU y China, más que las de empresas, pues no centramos en la geopolítica de tales pasos. Además, haremos hincapié en algunas de grupos profesionales, como el IEEE.

Tampoco entraremos aquí en algo que toca, sin embargo, plenamente a la ética de la IA, como son sus aplicaciones militares, que dejamos para otra ocasión. Aunque coincidimos con Mark Coeckelbergh en que muchas cosas que se aplican a la IA civil pueden cambiar el ámbito militar. Por ejemplo, la idea de redes o enjambres de robots, en las que no hay una inteligencia central que controle, sino distribuida, que plantea problemas de cambio de la dirección de la guerra y de responsabilidad.[20]

Se plantea también la cuestión de si los robots son meros medios, o pueden convertirse en fines, o al menos contribuir a conformar esos fines. Es decir, si hay que pensar en máquinas con un "estatus moral",[21] no tanto en el sentido de que sientan, sino de que sepan. Sapiencia antes que sensibilidad. Sobre todo, porque muchas maquinas están y estarán aún más, conectadas entre sí e intercambiando constantemente información en un grado muy superior a las personas. En este sentido, no serán personas.

Connected Driving (Ethical rules for self-driving computers) (2017), https://www.bmvi.de/SharedDocs/EN/publications/action-plan-on-the-report-ethics-commission-acd.pdf?__blob=publicationFile.

[20] Mark Coeckelbergh (2011), "From killer machines to doctrines and swarms, or why ethics of military robotics is not (necessarily) about robots", *Philosophy & Technology*, September, vol. 24, nº 3, pp. 269-278, https://doi.org/10.1007/s13347-011-0019-6.

[21] Nick Bostrom y Eliezer Yudkowsky (2011), "The ethics of Artificial Intelligence", https://nickbostrom.com/ethics/artificial-intelligence.pdf.

Diferencias político-culturales y principios comunes

Hay diferencias culturales y políticas en la manera cómo distintos países y empresas abordan la cuestión de la ética y la IA, aunque muchas veces no se hagan explícitas. Pocos de los planteamientos abordan este factor. El IEEE es una excepción en estos planteamientos.

Se plantea el problema de que los sistemas afectivos, es decir, la IA afectiva o emociones sintéticas,[22] que están desarrollándose para aumentar la accesibilidad a la IA, cuando se insertan a través de distintas culturas pueden afectar negativamente los valores culturales/sociales/religiosos de las comunidades y engañar a los humanos.
También hay que considerar que distintas culturas pueden tener diferentes enfoques de la ética en la IA, en lo que entran pocas propuestas.[23] En Japón, por ejemplo, pesa la cultura religiosa del sintoísmo respecto a los seres y a las cosas, como hemos mencionado en otra parte,[24] y como ha estudiado la antropóloga estadounidense Jennifer Robertson,[25] el sintoísmo atribuye características anímicas a muchas cosas; le preocupa ante todo la pureza y la polución y ve energías vitales (*kami*) en numerosos aspectos del mundo, ya sean árboles, rocas, personas, etc. Este enfoque facilita o hace más natural la relación con las máquinas, y los robots lo son, de forma muy especial. Y hay otras tradiciones como la Ubuntu en África, que señala que "una persona es una persona a través de otra gente" y de la pertenencia a comunidades.[26]

Los regímenes políticos liberales y los autoritarios tienen diferentes enfoques, aunque hay terrenos para posibles entendimientos. El caso de China (que se explica en el epígrafe correspondiente) es el más significativo, dada su potencia económica y demográfica y su peso en el desarrollo de la IA. Pero también hay diferencias entre el enfoque estadounidense y el europeo.

Pak-Hang Wong,[27] del departamento de Informática de la Universidad de Hamburgo, planteó la cuestión de la "innovación responsable para pueblos no liberales", para cuestionar el monopolio de la ética por parte de las tradiciones éticas occidentales. Rechaza la idea de que "los pueblos no liberales sean incapaces de investigar e innovar responsablemente porque no tienen el conjunto adecuado de valores para fundamentar los objetivos de la investigación e innovación, ni pueden justificar la innovación responsable con sus fundamentos normativos". Cita, por ejemplo, la idea de *Minben* ("el pueblo como base") en la filosofía política confuciana, que pone un énfasis en la responsabilidad del gobierno (o del gobernante) por el bienestar de la gente, aunque

[22] Un ejemplo de proyecto de investigación en India sobre este tema está en http://www.iitg.ac.in/cse/robotics/?page_id=392.
[23] La del IEEE, de la que partimos y que ampliaremos más adelante, es una excepción.

[24] Andrés Ortega (2016), *La imparable marcha de los robots*, Alianza Editorial.

[25] Jennifer Robertson (2007), "Robo Sapiens Japanicus: humanoid robots and the posthuman family", *Critical Asian Studies*, vol. 39, nº 3; y Jennifer Robertson (2010), "Robots of the rising sun", *The American Interest*, otoño (septiembre/octubre).

[26] Véase a este respecto IEEE.

[27] Pak-Hang Wong (2016), "Responsible innovation for decent nonliberal peoples: a dilemma?", *Journal of Responsible Innovation*, vol. 3, nº 2, pp. 154-168, DOI:10.1080/23299460.2016.1216709, http://dx.doi.org/10.1080/23299460.2016.1216709.

desde un enfoque paternalista. En segundo lugar, ve moralmente problemático la imposición de valores desde el exterior.

Para Wong, el primer desafío es teórico, es decir, la inclusión de otras culturas, lo que invita al relativismo (o a un pluralismo impotente) en el sentido de que no está claro qué valores incluidos en la innovación responsable, y, por tanto, cabe añadir en la IA, y qué valor(es), si existe (n), debe(n) tener prioridad en caso de conflicto. El segundo desafío es práctico, es decir, hay poca investigación sobre bases reguladoras alternativas para la innovación responsable y también para la IA, aunque se está progresando.

A pesar de las diferencias, muchas de las propuestas giran en torno a unos principios o conceptos comunes para la IA:

- Centrada en los humanos (*human centric*).
- Confiabilidad de la IA (*trustworthy*).
- Respeto a la autonomía humana.
- Prevención de hacer daño.
- Equidad y que "nadie se quede atrás".
- Explicabilidad.

(4) La visión desde los valores de la UE

Diversas instituciones de la UE han estado o están haciendo propuestas en materia de ética de la IA, desde la Comisión Europea al Parlamento Europeo pasando por el Grupo de Expertos de Alto Nivel sobre IA.[28] Cuenta también con una Alianza, civil, para la IA. No se trata sólo de liderar en los contenidos tecnológicos de la IA, sino también de lograrlo en materia de regulación, lo que puede tener un alcance global. La UE tiene ya un recorrido, por ejemplo, con el reglamento GDPR[29] para proteger la privacidad y voluntariedad de los datos personales. Su narrativa, centrada en los humanos, incide en proteger a los humanos y a la sociedad.

Según esta visión europea, la IA confiable debe ser:

(a) Legal: respetando todas las leyes y regulaciones aplicables.
(b) Ética: respetando los principios y valores éticos.
(c) Robusta: tanto desde una perspectiva técnica como teniendo en cuenta su entorno social.

Las directrices presentan un conjunto de siete requisitos clave que los sistemas de inteligencia artificial deben cumplir para ser considerados confiables:[30]

[28] "Piloting phase of ethical guidelines", https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai y https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top.

[29] Versión española en https://www.boe.es/doue/2016/119/L00001-00088.pdf.

[30] Añadimos consideraciones de otros documentos de la UE.

(a) Agencia y supervisión humana: los sistemas de inteligencia artificial deberían empoderar a los seres humanos, permitiéndoles tomar decisiones informadas y fomentar sus derechos fundamentales. Al mismo tiempo, se deben garantizar mecanismos de supervisión adecuados. La IA debe de ser supervisada por seres humanos, con las "apropiadas medidas de contingencia".

(b) Robustez técnica y seguridad: los sistemas de IA deben ser resistentes, resilientes y seguros ante eventuales intentos de manipulaciones o de pirateo y dotarse de planes de contingencia. Deben ser seguros, garantizar un plan alternativo en caso de que algo salga mal, además de ser precisos, confiables y reproducibles. Esa es la única forma de garantizar que también se puedan minimizar y prevenir daños no intencionales.

(c) Privacidad y gobernanza de datos: se debe garantizar la privacidad de los datos de los ciudadanos en todo el ciclo vital de la IA. Además de garantizar el pleno respeto de la privacidad y la protección de datos, también deben garantizarse mecanismos adecuados de gobernanza de datos, teniendo en cuenta la calidad e integridad de los datos, y garantizando el acceso legítimo a los datos.

(d) Transparencia: la IA debe de ser transparente, lo que supone poder reconstruir cómo y por qué se comporta de una determinada manera y quienes interactúen con esos sistemas deben de saber que se trata de IA así como qué personas son sus responsables. Los modelos de negocio de datos, sistema e IA deben ser transparentes. Los mecanismos de trazabilidad pueden ayudar a lograr esto. Además, los sistemas de IA y sus decisiones deben explicarse de manera adaptada a las partes interesadas en cuestión. Los seres humanos deben ser conscientes de que están interactuando con un sistema de IA y deben estar informados de las capacidades y limitaciones del sistema.

(e) Diversidad, no discriminación y equidad: la IA debe de tener en cuenta la diversidad social desde su desarrollo para garantizar que los algoritmos en que se base no tengan sesgos discriminatorios directos o indirectos. Se debe evitar el sesgo injusto, ya que podría tener múltiples implicaciones negativas, desde la marginación de los grupos vulnerables hasta la exacerbación de los prejuicios y la discriminación. Al fomentar la diversidad, los sistemas de IA deben ser accesibles para todos, independientemente de cualquier discapacidad, e involucrar a las partes interesadas relevantes a lo largo de todo su círculo de vida.

(f) Bienestar social y ambiental: el desarrollo tecnológico debe de tener en cuenta su impacto social y medioambiental de forma que sea sostenible y ecológicamente responsable. Los sistemas de IA deberían beneficiar a todos los seres humanos, incluidas las generaciones futuras. Por lo tanto, debe garantizarse que sean sostenibles y respetuosos con el medio ambiente. Además, deben tener en cuenta el medio ambiente, incluidos otros seres vivos, y su impacto social y social debe considerarse cuidadosamente.

(g) Responsabilidad: la IA y sus resultados deben de rendir cuentas ante auditores externos e internos. Se deben establecer mecanismos para garantizar la

responsabilidad y la responsabilidad de los sistemas de IA y sus resultados. La auditabilidad, que permite la evaluación de algoritmos, datos y procesos de diseño, desempeña un papel clave, especialmente en aplicaciones críticas. Además, debe garantizarse una reparación adecuada y accesible de la IA y los robots.

La Comisión Europea ha presentado estas guías éticas a los Estados miembros de la UE y a los distintos actores del sector y diseñado una fase piloto (hasta el pasado 1 de diciembre) para obtener retroalimentación de quienes están implicados en el salto tecnológico que suponen las máquinas capaces de aprender y decidir por sí mismas.

A partir de esa información, se actualizarán las guías sobre ética robótica a inicios de 2020 de la Comisión Europea, que planea destinar 1.000 millones de euros anuales a partir de 2020 a la IA y espera que en total se movilicen 200.000 millones en la próxima década.

España está pendiente de publicar su Estrategia Nacional de Inteligencia Artificial, que debería incluir consideraciones y propuestas sobre ética como una de sus prioridades. En la *Estrategia española de i+d+i en inteligencia artificial*[31] se indica la intención de encargar a un Comité Español de Ética en la Investigación (CEEI) tratar los temas éticos del uso e implementación de la IA. Asimismo, se menciona el papel de España a nivel europeo sobre este tema para redactar un Código Ético de la IA.

Destacan otros movimientos entre Estados miembros de la UE y más allá. La Comisión de Ética de Datos, por encargo del Gobierno alemán, ha presentado en octubre de 2019 sus propias propuestas, u opinión,[32] para las que pide un marco europeo. No se separa marcadamente de la línea general de la Comisión Europea. Insiste en que las consideraciones éticas se aborden "a lo largo de todo el proceso de desarrollo y aplicación de la IA", utilizando el enfoque "ética por, en y para el diseño" y como marca registrada la de *AI Made in Europe*. Considera que se necesita un amplio abanico de mecanismos de control para insertar los principios éticos y jurídicos en el proceso del diseño y la aplicación de estas tecnologías. Estos mecanismos deben decidirse a nivel nacional y europeo en un proceso democrático.

Los principios que reclama son los de dignidad humana, autodeterminación incluido la informacional para el control de los datos personales, el derecho a la privacidad, la seguridad, la democracia, la justicia y la solidaridad, y la sostenibilidad económica, ecológica y social. También estima necesario considerar las interacciones entre la tecnología, los usuarios y la sociedad ("el ecosistema de IA"). Y dentro de este ecosistema, garantizar la transparencia suficiente, la rendición de cuentas, la libertad frente a la discriminación y la capacidad de revisar los procesos automatizados que

[31] Estrategia Española de I+D+I en Inteligencia Artificial (2019), http://www.ciencia.gob.es/stfls/MICINN/Ciencia/Ficheros/Estrategia_Inteligencia_Artificial_IDI.pdf.

[32] The Federal Government, Data Ethics Commission, https://www.bmjv.de/DE/Themen/FokusThemen/Datenethikkommission/Datenethikkommission_EN_node.html; The Federal Government, Opinion of the Data Ethics Commission, https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=1.

preparan decisiones o extraer conclusiones que pueden llevarse a cabo sin aporte humano. Esta "es la única manera de generar confianza en el uso y los resultados de los procesos controlados por algoritmos".

En agosto de 2019, Alemania, Francia y Japón se han unido para financiar la investigación de inteligencia artificial "centrada en el ser humano" que tiene como objetivo respetar la privacidad y la transparencia, en lo que algunos observadores califican como última señal de una división global con EEUU y China sobre la ética de la IA. Las agencias de financiación de los tres países han presentado una convocatoria conjunta para propuestas de investigación, respaldadas por 7,4 millones de euros iniciales (8,2 millones de dólares) desde un punto de partida: "compartir los mismos valores" ante el potencial de la tecnología de "violar la privacidad individual y el derecho a la autodeterminación informativa".[33]

Otras propuestas han emanado del ámbito europeo. El Consejo de Europa se ha centrado en la protección de datos al publicar su Comité Consultivo para la Protección de Individuos en Relación al Procesamiento de Datos Personales unas Líneas Guía sobre Inteligencia Artificial y Protección de Datos Personales,[34] cuyo objeto es ayudar a los decisores políticos, los desarrolladores de IA, los fabricantes y los proveedores de servicios a asegurar que las aplicaciones de IA no socaven el derecho a la protección de datos. Subraya que la protección de los derechos humanos, incluido el derecho a la protección de datos personales, debe ser un requisito previo esencial al desarrollar o adoptar aplicaciones de IA, en particular cuando se utilizan en los procesos de toma de decisiones, y debe basarse en los principios del convenio actualizado de protección de datos, el Convenio 108+,[35] abierto a la firma en octubre de 2018. La UNESCO también ha trabajado en el tema y la ISO (Organización Internacional de Normalización) también está desarrollando estándares éticos para la IA.[36]

(5) OCDE Y G20

Quizá por su amplitud geográfica y económica, una de las iniciativas más interesantes es la de la OCDE. Cuarenta y dos 42 países (los 36 de la OCDE, incluido EEUU, más Argentina, Brasil, Colombia, Costa Rica, Perú y Rumanía) firmaron en mayo de 2019 unos Principios sobre IA.[37] [38] En resumen, proponen que:

[33] David Matthews (2019), "New Research Alliance Cements Split on AI Ethics", *Inside Higher ED*, 23/VIII/2019, https://www.insidehighered.com/news/2019/08/23/new-research-alliance-cements-split-ai-ethics.

[34] "Guidelines on Artificial Intelligence and Data Protection", enero 2019, https://rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8; y "European Ethical Charter on the Use of AI in Judicial Systems and their Environment", https://www.coe.int/en/web/cepej/cepej-european-ethical-charter-on-the-use-of-artificial-intelligence-ai-in-judicial-systems-and-their-environment.

[35] Council of Europe, "Convention 108 and Protocols", https://www.coe.int/en/web/data-protection/convention108-and-protocol.

[36] Robert Bartram (2018), "The new frontier for artificial intelligence", *International Organization for Standardization*, 18/X/2018, https://www.iso.org/news/ref2336.html.

[37] Council of Europe (2019), "Guidelines on Artificial Intelligence and Data Protection", https://rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8.

[38] OECD (2019), "Recommendation of the Council on Artificial Intelligence", https://legalinstruments.oecd.org/api/print?ids=648&lang=en; "Forty-two countries adopt new OECD
(cont.)

(a) La IA debería beneficiar a las personas y al planeta al impulsar el crecimiento inclusivo, el desarrollo sostenible y el bienestar.

(b) Los sistemas de IA deben diseñarse de manera que respeten el Estado de Derecho, los derechos humanos, los valores democráticos y la diversidad, y deben incluir salvaguardas apropiadas, por ejemplo, permitiendo la intervención humana cuando sea necesario, para garantizar una sociedad justa y equitativa.
(c) Debe haber transparencia y divulgación responsable en torno a los sistemas de IA para garantizar que las personas entiendan cuándo se relacionan con ellos y puedan cuestionar los resultados.

(d) Los sistemas de IA deben funcionar de manera robusta y segura a lo largo de sus vidas, y los riesgos potenciales deben evaluarse y gestionarse continuamente.

(e) Las organizaciones e individuos que desarrollan, implementan u operan sistemas de inteligencia artificial deben ser responsables de su correcto funcionamiento de acuerdo con los principios anteriores.

La OCDE recomienda a los gobiernos:

- Facilitar la inversión pública y privada en investigación y desarrollo para estimular la innovación en IA confiable.

- Fomentar ecosistemas de IA accesibles con infraestructura y tecnologías digitales, y mecanismos para compartir datos y conocimientos.

- Crear un entorno de políticas que abrirá el camino para la implementación de sistemas de IA confiables.

- Equipar a las personas con las habilidades para la IA y apoyar a los trabajadores para garantizar una transición justa.

- Cooperar a través de las fronteras y sectores para compartir información, desarrollar estándares y trabajar hacia la administración responsable de la IA.

En la Reunión Ministerial del G20 sobre Comercio y Economía Digital, celebrada del 8 al 9 de junio de 2019 en Tsubuka, Japón, los ministros aprobaron como anexo a su declaración unos principios para la IA,[39] ratificados en la posterior cumbre de Osaka, que, como reconocen, beben en la propuesta de la OCDE. Han sido firmados por todos los participantes incluidos regímenes tecno-autoritarios como China y Rusia y otros como Arabia Saudí.

Principles on Artificial Intelligence", https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm.

[39] Ministry of Economy, Trade and Industry of Japan (2019), "G20 Ministerial Statement on Trade and Digital Economy", https://www.meti.go.jp/press/2019/06/20190610010/20190610010-1.pdf.

En ellos se recogen los de crecimiento inclusivo, desarrollo sostenible y bienestar; valores centrados en el ser humano y equidad; transparencia y explicabilidad; robustez, seguridad y seguridad; y responsabilidad. También se propugna, junto a los impulsos nacionales, una cooperación internacional para una IA confiable.

Pero, hay que insistir, el problema no son los principios, sino su aplicación.

(6) EEUU: IA con valores norteamericanos

En la línea que abrió la Administración Obama, el Gobierno de Donald Trump, a través de una Orden Ejecutiva y de una Iniciativa para una IA Americana,[40] lanzó en febrero de 2019 la idea de una IA con valores norteamericanos, aunque no está tan desarrollada como la propuesta europea.

Según esta visión, las tecnologías IA deben reflejar "valores americanos fundamentales" como son la libertad, y garantizar los derechos humanos, el Estado de Derecho, la estabilidad de sus instituciones, los derechos a la privacidad, el respeto por la propiedad intelectual y las oportunidades "para todo el mundo de perseguir sus sueños". Además de reflejar la "devoción americana para ayudar a la gente".

Más en concreto, esta tecnología debe ser comprensible, digna de confianza y segura. También propugna examinar los efectos en la fuerza laboral y que la IA se desarrolle de forma responsable –lo que, por su parte, está desarrollando la National Science Foundation (NSF), con un programa específico para comprender mejor "la asociación entre la tecnología humana y el panorama sociotecnológico emergente, para crear nuevas tecnologías que aumenten el rendimiento humano y fomenten el aprendizaje vivo y generalizado con la tecnología"–.[41]

Las agencias federales impulsarán la confianza pública en la IA estableciendo guías (*guidance*) para el desarrollo y uso de la IA por diversos actores.

Si se insiste desde Washington en la necesidad de explicabilidad de lo que hace la IA, y a este respecto, la agencia DARPA ligada al Pentágono tiene un programa XAI sobre explicabilidad de la IA[42] –DARPA's Explainable AI (XAI) program–. Por su parte, la NSF está desarrollando con Amazon un programa sobre *fairness* (justicia) en la IA.[43]

[40] The White House, "AI with American values", https://www.whitehouse.gov/ai/ai-american-values/ y "Artificial Intelligence for the American people", https://www.whitehouse.gov/ai/.

[41] Future of Work at the Human-Technology Frontier, https://www.nsf.gov/eng/futureofwork.jsp.

[42] DARPA's Explainable AI (XAI) Program, https://www.darpa.mil/program/explainable-artificial-intelligence.

[43] NSF Program on Fairness in Artificial Intelligence (AI) in Collaboration with Amazon (FAI), https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505651.

(7) Tecno-utilitarismo/autoritarismo: China y Rusia

China ha tardado, pero finalmente también ha entrado en la cuestión de la ética de la IA, en lo que ha llamado los "principios de Pekín de IA".[44] Son 15 puntos elaborados por la Academia de Pekín de Inteligencia Artificial (BAAI), apoyada por el Ministerio de Ciencia y Tecnología y el ayuntamiento de la capital, en colaboración con las principales centros y empresas de IA en China, de la Universidad de Pekín a las empresas Baidu, Alibaba y Tencent. Se han hecho públicas a mediados de 2019, cuando las tensiones en materia de tecnologías y comercio entre EEUU y China estaban en escalonado.[45]

China diferencia entre los principios para la investigación y el desarrollo de la IA y los que se aplican a su uso. Entre los primeros: hacer el bien, servir a la humanidad –que implica entre otras cosas que "la privacidad, la dignidad, la libertad, la autonomía y los derechos humanos deben ser suficientemente respetados". Propugna que la IA no dañe a los humanos, sea responsable, controlar los riesgos, ser diversa e inclusiva, y ser abierta y compartida. En cuanto a los del uso, cita la necesidad de aplicarla "sabia y adecuadamente, con consentimiento informado, con educación y formación".

A diferencia de la edición genética, opina Chen Xiaoping, que encabezó el comité de Ética que redacto este texto, el riesgo está en la aplicación no en la tecnología en sí misma.[46] Además, añade otros seis principios en cuestión de gobernanza (optimizar el empleo, armonía y cooperación para lograr una "simbiosis óptima", adaptación y moderación, subdivisión e implementación, y planeamiento a largo plazo).

En materia de "ser ética", la I+D debe enfocar el enfoque de la IA para hacerla confiable (*trustworthy*), entre lo que incluye hacer el sistema lo más justo posible, reducir los posibles discriminaciones y sesgos, y hacer el sistema más trazable, auditable y que responda (*accountable*). Hace, asimismo, un llamamiento a la cooperación internacional.

Tencent habla de un "contrato social" entre empresas y usuarios que regule el uso de datos, de "tecnología para el bien". Tencent creó en 2018 una plataforma de IA para el Bien Social un año antes que Google diera un paso parecido. Su fundador y CEO, Pony Ma, ha propuesto un marco ético para la gobernanza de la IA, en torno a cuatro principios: disponibilidad, fiabilidad, totalidad y controlabilidad. Cree necesario avanzar en incluir estas consideraciones éticas en el diseño de la IA.[47]

Estas propuestas indican que China, hacia fuera y en algunos círculos internos, está dispuesta a debatir estos temas. Brinda la posibilidad de tender puentes en el debate, pero no cabe olvidar que se ha constituido en un Estado de Vigilancia Digital, desde una ideología totalitaria. Pero, dada la importancia de China en el desarrollo de la IA, puede

[44] Beijing AI principles, 28/V/2019, https://www.baai.ac.cn/blog/beijing-ai-principles.

[45] Will Knight (2019), "Why does Beijing suddenly care about AI ethics?", *MIT Technology Review*, 31/V/2019.

[46] "China's top AI scientist drives development of ethical guidelines", *South China Morning Post*, 10/I/2019.

[47] Jason Si, Dean of Tencent Research Institute (2019), "These rules could save humanity from the threat of rogue AI", World Economic Forum, 8/V/2019, https://www.weforum.org/agenda/2019/05/these-rules-could-save-humanity-from-the-threat-of-rogue-ai/; y *Financial Times*, 1/V/2019.

influir en este terreno. Kai-Fu Lee, en un libro que ha tenido un gran eco,[48] ha señalado que este país puede liderar en muchos campos de aplicación de la IA, y tiene un enfoque "tecno-utilitario". Lee incide en que China prioriza la idea del mayor bien para el mayor número en lugar de un imperativo moral para proteger los derechos individuales, que predomina en Occidente. Los consumidores chinos parecen menos preocupados por la privacidad, incluidos los sistemas de reconocimiento facial para muchas aplicaciones, además de funciones de vigilancia. Hacen otro tipo de *trade-off*, de compensación, entre la vigilancia y la conveniencia, como bien lo expone John Thornhill.[49] Esa, sugiere Lee, es una de las razones por las que los consumidores chinos están menos preocupados por la instalación de dispositivos de reconocimiento facial en carritos de supermercado, para personalizar los viajes de compras, o en las aulas para detectar a los estudiantes inatentos.

En todo caso, China ha apoyado los principios del G20. Como veremos, el problema no son tanto las diferencias en los principios, como en su aplicación práctica.

Por su parte, el presidente de la Federación Rusa, Vladimir Putin, aprobó un decreto en octubre de 2019 con las líneas maestras de la Estrategia Nacional para el Desarrollo de la Inteligencia Artificial para el periodo 2020-2030.[50] Algunos de sus aspectos ya estaban contenidos en el borrador que Putin encargó a Sherbank.[51] El año pasado ya se publicó una hoja de ruta, con 10 puntos, ninguno de los cuales abordaba directamente las cuestiones éticas,[52] salvo determinar "la distribución de la responsabilidad entre propietarios, desarrolladores y proveedores de datos por daños causados mediante sistemas de IA", y aclarar "la regulación de la circulación de los resultados de la actividad intelectual utilizando la IA". Tampoco esta estrategia entra en asuntos éticos. Eso sí, determina la "creación de un sistema integrado para regular la acción" de la IA, que "formule reglas éticas para la interacción humana con la Inteligencia artificial". Advierte que "una regulación excesiva en esta esfera podría reducir significativamente el ritmo de desarrollo y la introducción de soluciones tecnológicas". Y establece que la protección de los datos obtenidos requerirá su almacenamiento "dentro de la Federación Rusa", así como el acceso prioritario a estos datos de las autoridades y organizaciones públicas rusas.

[48] Kai-Fu Lee (2018), *AI Superpowers: China, Silicon Valley, and the New World Order*, Houghton Mifflin Harcourt.
[49] John Thornhill (2019), "Formulating values for AI is hard when humans do not agree", *Financial Times*, 22/VII/2019, https://www.ft.com/content/6c8854de-ac59-11e9-8030-530adfa879c2.
[50] "Russian president approves national AI strategy", 10/X/2019, https://dig.watch/updates/russian-president-approves-national-ai-strategy. El texto oficial en ruso está disponible en http://namib.online/wp-content/uploads/2019/10/%D0%9E-%D1%80%D0%B0%D0%B7%D0%B2%D0%B8%D1%82%D0%B8%D0%B8-%D0%B8%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE-%D0%B8%D0%BD%D1%82%D0%B5%D0%BB%D0%BB%D0%B5%D0%BA%D1%82%D0%B0-%D0%B2-%D0%A0%D0%A4.pdf. Traducción del CSET en https://cset.georgetown.edu/wp-content/uploads/t0060_Russia_AI_strategy_EN-1.pdf.
[51] Samuel Bendett (2019), "First Draft of Russia's AI Strategy", *Defense One*, 10/IX/2019, https://www.defenseone.com/technology/2019/09/whats-russias-national-ai-strategy/159740/.
[52] Samuel Bendett (2018), "Here's How the Russian Military Is Organizing to Develop AI", *Defense One*, 20/VII/2018, https://www.defenseone.com/ideas/2018/07/russian-militarys-ai-development-roadmap/149900/?oref=d-river.

Rusia también ha suscrito los principios del G20.

(8) La visión desde la ingeniería

La ingeniería de *software*, y la IA lo es en gran parte, no se ha reconocido como una profesión con deberes fiduciarios hacia el público. Pero a este respecto, está cambiando.

Diversos estudios recientes señalan que los códigos éticos establecidos han tenido escaso efecto o impacto en las decisiones cotidianas de los profesionales y estudiantes de ingeniería de *software*.[53] Pero, como hemos indicado, empiezan a multiplicarse los cursos de ética para ingenieros. A la vez, estos enriquecen el debate.

(8.1) Ingenieros éticos

Robert McGinn, en su libro *The Ethical Engineer. Contemporary Concepts & Cases*[54], observa una desconexión entre la formación ética de los estudiantes de ingeniería (lado teórico) y la realidad del trabajo de ingeniería actual que cada día, gracias a la tecnología, es más ubicuo (lado practico).

Según este autor, los ingenieros tienen que asumir –como parte de su profesión– un único deber fundamental, a saber, combatir el daño. Para ello, plantea cuatro responsabilidades éticas o FERE (*Fundamental Ethical Responsibility of Engineers*):

(a) No causar daño o no crear un irrazonable riesgo de daño a los demás (ni tampoco al bienestar e intereses públicos) con su trabajo (FERE1).

(b) Intentar prevenir el daño o cualquier riesgo irrazonable de daño a los demás (ni tampoco al bienestar e intereses públicos) que pueda ser causado por su trabajo como ingeniero, por el trabajo de ingeniería de otras personas en el que el ingeniero está involucrado, o por un trabajo que dicho ingeniero sea conocedor (FERE2).

(c) Intentar alertar e informar sobre el riesgo del daño a las personas y a los colectivos que puedan sufrir un riesgo irrazonable y verse afectadas por su trabajo de ingeniería: por el trabajo de ingeniería de otras personas donde el ingeniero está involucrado o por un trabajo del que dicho ingeniero sea conocedor (FERE3).

(d) Trabajar lo mejor posible para atender los intereses legítimos de su empleador o cliente (FERE4 que se aplica en la mayoría de los casos a los ingenieros empleados en una empresa o contratados por un cliente. Incluso si el ingeniero no tiene relaciones laborales o contractuales con otras personas, se puede decir que el

[53] Entre ellos: A. McNamara, J. Smith y E. Murphy-Hill (2018), "Does ACM's code of ethics change ethical decision making in software development?", *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering – ESEC/FSE 2018*, ACM Press, doi:10.1145/3236024.3264833.

[54] McGinn, Robert (2018), *The Ethical Engineer. Contemporary Concepts & Cases*, Princeton University Press. Agradezco a Migle Laukyte, de la Universidad Carlos III de Madrid, sus comentarios al respecto recogidos en el *Observatorio de las Ideas*, nº 64-65, julio-agosto de 2018.

ingeniero es "empleado" por la sociedad misma que le permite de practicar su profesión).

Las cuatro responsabilidades éticas no son suficientes para convertir un ingeniero en un ingeniero ético. Hay cuatro factores más para llegar a este objetivo:

(a) Factores sociotécnicos del trabajo de ingenieros, como por ejemplo si el ingeniero trabaja en solitario o en equipo.

(b) El contexto socio-organizativo del lugar de trabajo del ingeniero, por ejemplo, si la empresa del ingeniero está bajo presión económica y cuál es la cultura laboral en la empresa.

(c) El contexto macrosocial en el que los resultados del trabajo de ingeniero van a ser utilizados. Hay mucha diferencia entre un mercado en el que el consumidor conoce sus derechos y está bien protegido, y otro en el que la cultura de protección de consumidor está solo empezando a germinar.

(d) Las consecuencias dañosas y beneficiosas de dichos resultados del trabajo: los ingenieros tienen que preguntarse a quién aporta beneficios su trabajo y a quién, al contrario, daña.

(8.2) El enfoque holístico del IEEE

El Instituto de Ingeniería Eléctrica y Electrónica (IEEE en sus siglas en inglés) ha agrupado para su iniciativa colaborativa a varios centenares de líderes de seis continentes del mundo académico, de la industria, de la sociedad civil y de la administración de todo el mundo para presentar una propuesta, *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*,[55] en su publicación, en versión definitiva tras varios borradores ampliamente discutidos. Su *Ethically Aligned Design* (diseño éticamente alineado) resulta ser de las propuestas más ricas y completas de las aquí analizadas, porque entra en la necesidad de estándares ético-técnicos –propone 14– y en la necesidad de autoridades de certificación. Es de las pocas organizaciones que se orientan a incrustar las normas y valores humanos en la IA.

De hecho, rechazan el termino IA para conceptualizar y hablar de tecnologías y sistemas que extienden nuestra inteligencia humana o se usan en aplicaciones robóticas. Prefieren el término "sistemas autónomos e inteligentes". Aunque por claridad y comparabilidad con otras propuestas aquí seguimos utilizando el termino IA.

El IEEE plantea "principios generales como imperativos", no muy diferentes de los que ya hemos venido viendo en otras propuestas: derechos humanos, bienestar humano,

[55] "The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2", 2018, https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf. La versión definitiva se encuentra en https://ethicsinaction.ieee.org/.

agencia de datos para mantener la capacidad de las personas para tener control sobre su identidad, eficacia, transparencia, responsabilidad, concienciación sobre el uso indebido y riesgos de la IA, y competencia y habilidad necesaria para un funcionamiento seguro y eficaz de la IA.

Se proponen tres pilares para el marco conceptual del "diseño éticamente alineado":

(a) Valores Humanos Universales: la IA puede ser una "enorme fuerza para el bien en la sociedad" siempre que respete los derechos humanos, se aliena con los valores humanos y aumente holísticamente el bienestar al tiempo que empodera a tantas personas como sea posible. También debe diseñarse para salvaguardar el medio ambiente y los recursos naturales.

(b) Una Agencia Política de Autodeterminación y Datos: si está correctamente diseñada e implementada, tiene un gran potencial para fomentar la libertad política y la democracia, mejorar la eficacia del gobierno y su responsabilidad, y fomentar la confianza y proteger nuestra esfera privada, pero sólo cuando la gente tiene su identidad digital y sus datos están protegidos de manera probada.

(c) Confiabilidad técnica: la IA debe ofrecer servicios en los que se pueda confiar. Las tecnologías deben ser monitoreadas para asegurar que su funcionamiento cumpla con objetivos éticos predeterminados. Los procesos de validación y verificación deben conducir a una mejor auditabilidad y certificación.

El objetivo del IEEE es "aportar ideas pragmáticas y direccionales y recomendaciones, que sirven de referencia clave para la labor de los tecnólogos, los educadores y los encargados de la formulación de políticas en los próximos años", y "avanzar en una discusión pública sobre cómo podemos establecer implementaciones éticas y sociales en sistemas y tecnologías inteligentes y autónomos, alineándolos con valores definidos y principios éticos que priorizan el bienestar humano en un contexto cultural determinado". Para ello, proponen inspirar la creación de Estándares (IEEE P7000™ serie y más allá) y programas de certificación, y facilitar el surgimiento de políticas nacionales y globales que se alineen con estos principios y resulten beneficiosos para las personas más allá de alcanzar los objetivos funcionales y problemas técnicos.

Busca asegurar que todas las partes interesadas involucradas en el diseño y desarrollo de sistemas autónomos e inteligentes estén educados, capacitados y empoderados para priorizar las consideraciones éticas a fin de que estas tecnologías sean avanzadas para el beneficio de la humanidad.

Pero el informe del IEEE alerta que introducir normas en dichos sistemas "requiere una clara delineación de la comunidad en los que se van a desplegar", lo que a su vez ha de tener en cuenta los aspectos culturales, que hemos mencionado.

Para abordar mejor las cuestiones de responsabilidad y responsabilidad, propone:

(a) Que los parlamentos/tribunales aclaren las cuestiones de responsabilidad, culpabilidad, responsabilidad y responsabilidad por lo que llamamos IA siempre que sea posible durante la fase de desarrollo e implementación (para que sus fabricantes y usuarios entiendan su derechos y obligaciones).

(b) Los diseñadores y desarrolladores de IA deben mantener el conocimiento de, y tener en cuenta cuando es relevante, la diversidad de la cultura existente, para las normas entre los grupos de usuarios de estos sistemas.

(c) Los ecosistemas de múltiples partes interesadas deberían ayudar a crear normas (que pueden madurar las mejores prácticas y leyes) donde no existen.

El IEEE sugiere proporcionar educación ética y conciencia de seguridad para sensibilizar a la sociedad a los riesgos potenciales de uso indebido de la IA (por ejemplo, proporcionando advertencias de "privacidad de datos" que algunos dispositivos inteligentes recogen los datos personales). Y hacerlo de forma escalable y efectiva, empezando por los que gozan de mayor credibilidad e impacto (instituciones, etc.), utilizando para ello también las redes sociales. También hay que educar a los funcionarios, políticos, legisladores y a las agencias responsables de la aplicación de estos principios, de modo que trabajen con los ciudadanos de forma colaborativa (incluido en los colegios).

El informe del IEEE alerta de que los comités de ética no tienen los recursos suficientes, que hay que generar un nuevo acceso al vocabulario clásico de la ética tanto para los ciudadanos, como para las empresas y los ingenieros, que diseñar para la seguridad puede ser mucho más difícil más tarde en el ciclo de vida del diseño en lugar de que antes, y que esta ética debe estar basada en reglas. Y que detrás de esto hay una cuestión de consentimiento y conciencia (*consent* y *awareness*). También señala que "es muy posible que hayamos llegado un punto en el que la IA está afectando a los humanos psicológicamente más de lo que creemos".

El objetivo sigue siendo el bienestar (*wellbeing*) humano. El bienestar, para el propósito de la Iniciativa Global de la IEEE, "abarca la satisfacción con la vida y las condiciones de vida como el equilibrio adecuado entre los efectos positivos y los afectos negativos" de la IA, en línea con la definición de la OECD.

Pero indica que la sociedad no ha establecido normas universales ni principios rectores para incorporar valores y normas en sistemas autónomos e inteligentes (A/IS) hoy en día. Abordar el objetivo general de la incorporación de normas y, por implicación, los valores en la IA, requiere, según este enfoque, abordar algunos objetivos más concretos, como identificar las normas de la comunidad en la que operan la IA, una implementación computacional de las normas de esa comunidad, y evaluar si la implementación de las normas identificadas en la IA es de hecho conforme a las normas reflexivas de esa comunidad. Para ello se requiere una investigación colaborativa entre científicos de diferentes escuelas de pensamiento y diferentes disciplinas.

(9) El conflictivo paso a la regulación

"Los principios no se traducen automáticamente en práctica", señala Mittelstadt.[56] Lo difícil es aplicarlos. Y no sólo por una cuestión técnica de incorporarlos a los algoritmos sino porque puede haber diversos enfoques entre países y entre empresas, especialmente porque todo ello requiere de regulación.

No se trata sólo de una cuestión de democracia liberales frente a regímenes iliberales o autoritarios. También vamos a ver –y estamos viendo– diferencias profundas en regulación por ejemplo entre EEUU y Europa, lo que puede llevar a divergencias y conflictos comerciales e incluso ideológicos. Por ejemplo, en los mecanismos para asegurar la transparencia o la privacidad. El uso de la IA es distinto entre EEUU y Europa, y con otros países, según se trate de gobiernos o empresas. Pero si EEUU y la UE no se ponen de acuerdo, China y otros países lo tendrán más fácil para imponer su regulación a una escala más global.

En todo caso está resultando difícil para cualquier Estado u organización regular el desarrollo del *software*, aunque la UE lo ha hecho para una parte de los datos con el GDPR, y ahora avanza en materia de IA. Y dados los avances en la tecnología, hay que tomarlo como un proceso, no como una solución definitiva, como también señala Mittlestadt. Bostrom[57] pide que la transparencia sea también a la inspección, una vez que hay regulación y que los inspectores estén capacitados para realizar sus funciones en este entorno.

Hay una gran resistencia por parte de algunas empresas a la regulación. Las empresas tienden más bien a propugnar una autoregulación, que, como se está comprobando en la práctica, no es suficiente. Es verdad que la realidad evoluciona más rápidamente que lo que pueden hacer los gobiernos. En 2011 el ex CEO de Google Eric Schmidt advirtió[58] que el exceso de intervención gubernamental limitaría la innovación: "Nos moveremos mucho más rápido que cualquier gobierno". Luego, en 2013, el cofundador de Google, Larry Page, se quejó de que "viejas instituciones como la ley" impiden la libertad de la compañía de "construir cosas realmente grandes".
Es quizá necesario idear un nuevo enfoque a la regulación. No se puede regular problemas presentes y futuros del siglo XXI con instrumentos, jurídicos y otros, del siglo XX. Para el IEEE hay que asegurar que la IA apoya, promueve y permite normas legales reconocidas internacionalmente.

Si es necesario un sustrato común global para esta regulación, pero no todos tienen que coincidir en todo. En todo caso, visto desde EEUU y Europa, si no hay un acuerdo básico transatlántico al respecto, se generará un vacío que llenarán otros por la vía del hecho o del derecho.

[56] Mittelstadt (2019), *op. cit.*

[57] Bostrom y Yudkovsky (2011), *op. cit.*

[58] Pascal-Emmanuel Gobry (2011), "Eric Schmidt To World Leaders At eG8: Don't Regulate Us, Or Else", *Business Insider*, 24/V/2011, https://www.businessinsider.com/eric-schmidt-google-eg8-2011-5?r=US&IR=T.

(10) Conclusiones

En realidad, se está maniobrando mucho en torno a la ética, pero pocos hablan de verdad de cómo se aplica la IA. La proliferación de *guidelines* no acerca a cómo implementarlos. Los ingenieros están más avanzados en sus reflexiones al respecto. Y la UE y algunos Estados miembros se están metiendo de lleno en ello, lo que les puede dar ventaja en términos de regulación. Hay diferencias entre regiones. Ya con el GDPR quedó patente el mayor valor que desde Europa se otorgaba a principios y valores como la privacidad o el derecho al olvido, aunque muchas empresas extraeuropeas se han visto obligadas a acatar este reglamento. Pero hemos de ser conscientes de que el GDPR reguló bien algunas cosas, pero poca gente lo ha sentido en su vida cotidiana.

Hay una carrera por la ética en IA que tiende fundamentos reales, pero que también encubre el deseo de algunos regímenes de cubrirse las espaldas ante las críticas. Hay elementos diferenciadores entre EEUU-Américas y Europa y China. EEUU tiene un enfoque más de riesgo, la UE más de privacidad y de principio de precaución, y China más de vigilancia estatal, sin privacidad.

Bajo la apariencia de consenso puede haber desacuerdos políticos y normativos profundo, lo que puede llevar a conflictos, no ya entre países, sino entre los gobiernos, las industrias y las sociedades civiles, como apunta Mittlestadt. Es una cuestión para abordar en las relaciones transatlánticas. Si EEUU (y en general las Américas) y Europa no se ponen de acuerdo en unas bases comunes de regulación ética práctica de la IA (y de otras materias), se puede generar un vacío que llenarán otros. No es necesario que el acuerdo sea total y exhaustivo. Y también ha de tener en cuenta a otros actores para ser realmente global. Pues la IA acabará por no conocer fronteras, aunque sí geopolítica.

La verdad llegará cuando estos principios éticos, sobre los que hay un amplio acuerdo, se introduzcan realmente en los algoritmos y sistemas. Y esa batalla por la reglamentación va a ser intensamente geopolítica. La regulación es poder.[59][60]

[59] Lorenzo Mariani y Micol Bertolini (2019), “The US-China 5G contest: options for Europe”, IAIA, http://www.tepsa.eu/the-winners-of-the-iai-essay-competition-discussed-in-rome-on-their-visions-of-europe-iai-italy-2-2-2-2-2-2/.

[60] Quisiera agradecer sus comentarios y aportaciones a Migle Laukyte, de la Universidad Carlos III de Madrid, y a Pablo Noriega, del Instituto de Investigación en Inteligencia Artificial (IIIA) del CSIC en Barcelona.