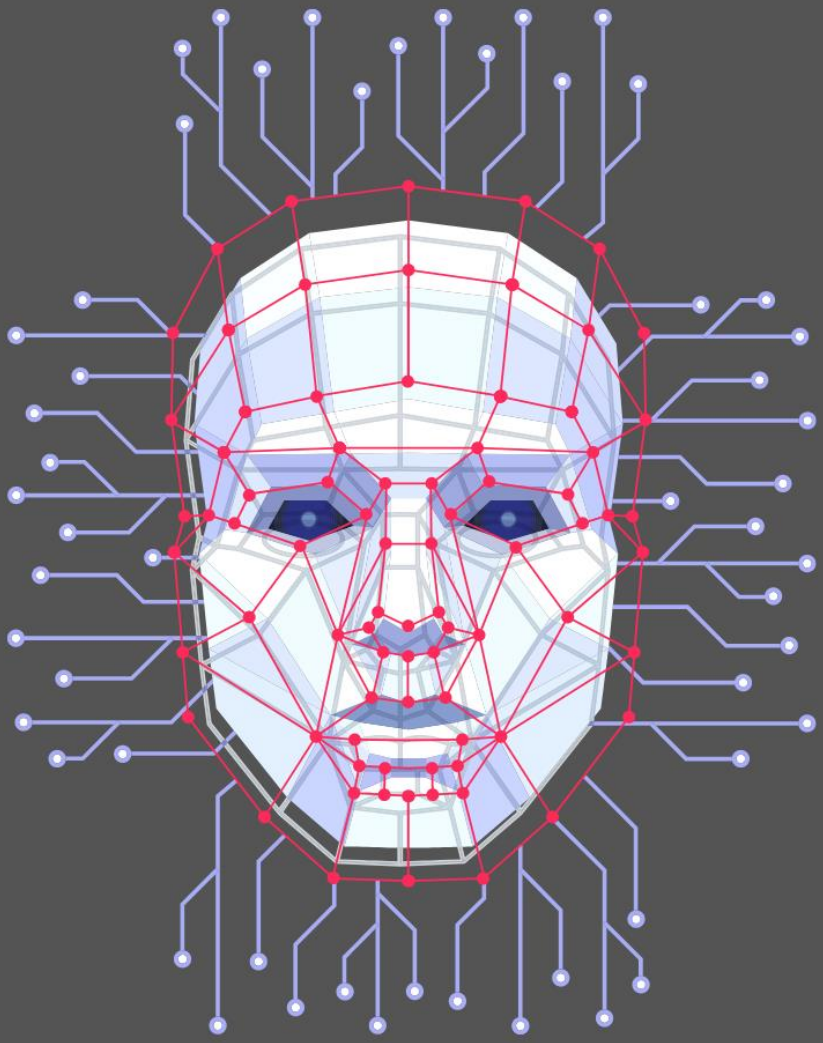
CONSEJO DE CIENCIA Y TECNOLOGÍA DEL ESTADO DE PUEBLA



PRINCIPIOS NORMATIVOS PARA UNA ÉTICA DE LA INTELIGENCIA ARTIFICIAL

FABIO MORANDÍN-AHUERMA

PUEBLA Un gobierno presente
Secretaría de Educación
CONCYTEP Consejo de Ciencia y Tecnología del Estado de Puebla
Benemérita Universidad Autónoma de Puebla 1578
Complejo Regional Nororiental

PRINCIPIOS NORMATIVOS PARA UNA ÉTICA DE LA INTELIGENCIA ARTIFICIAL

Fabio Morandín-Ahuerma

ISBN: 978-607-8901-78-4
Primera edición, México, 2023

IEEE: UN ESTÁNDAR GLOBAL COMO INICIATIVA ÉTICA DE LA IA

Introducción

"Diseño alineado éticamente: una visión para priorizar el bienestar humano con sistemas autónomos e inteligentes, primera edición" fue redactado bajo la "Iniciativa global sobre ética de los sistemas autónomos e inteligentes" y es un documento que pretende ofrecer un marco para las consideraciones éticas en el diseño, desarrollo, despliegue y uso de los sistemas de IA, con el objetivo de garantizar que las tecnologías digitales ayuden a las personas en sus labores. El Diseño consta de ocho principios: derechos humanos, bienestar, agencia de datos, eficacia, transparencia, rendición de cuentas, conciencia de uso indebido y competencia. Algunas de las observaciones a la iniciativa son que podría estar influida por los beneficios de la industria, y no represente las perspectivas e intereses de las comunidades digitalmente marginadas. A pesar de estas críticas, el documento del IEEE sigue siendo influyente en el campo de la ética y la gobernanza de la IA, y ha desempeñado un papel importante en la configuración del debate mundial en torno al desarrollo y despliegue responsables de la tecnología. También se analiza que el Instituto ha lanzado el "IEEE 7000™-2021, proceso de modelo estándar IEEE para abordar preocupaciones éticas durante el diseño del sistema" con el que sienta el precedente más objetivo de una ética de la IA aplicable.

La iniciativa global de IEEE

"Diseño alineado éticamente: una visión para priorizar el bienestar humano con sistemas autónomos e inteligentes, primera edición" (Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition) redactado bajo la "Iniciativa global sobre ética de los sistemas autónomos e inteligentes" (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems) [1] ha influido en el desarrollo de la ética y la gobernanza de los sistemas de IA, y ha sido adoptada por varias organizaciones y gobiernos del mundo como guía para el despliegue responsables de la IA.

El Instituto de ingenieros eléctricos y electrónicos (IEEE) es la mayor organización profesional técnica dedicada al avance de la tecnología en beneficio de la humanidad. Sus fundadores fueron, entre otros, Thomas Alva Edison y Alexander Graham Bell [2].

El IEEE cuenta con más de 423 000 miembros en más de 160 países, según su sitio web [3]. Entre sus miembros hay ingenieros, científicos y otros profesionales de la tecnología, así como estudiantes y educadores. La organización está conformada por varias sociedades, cada una de las cuales se centra en un área técnica específica, como la electricidad, la energía, las telecomunicaciones y, especialmente, la informática.

El IEEE puso en marcha la "Iniciativa global para la ética de los sistemas autónomos e inteligentes" (Global Initiative on Ethics of Autonomous and Intelligent Systems) [1] con el fin de abordar los problemas éticos relacionados con la creación y difusión de dichos sistemas. En la iniciativa se identificaron más de 120 cuestiones significativas y se sugirieron posibles soluciones [4]. Además, ha servido de base para 14 proyectos normalizados que están actualmente en curso a través de la Asociación de normas del IEEE. Los principios en el diseño éticamente alineado son:

1. Derechos humanos

> Los sistemas de IA deben ser creados y operados para respetar, promover y proteger los derechos humanos reconocidos internacionalmente *[1, p. 19].*

Para garantizar que el uso de sistemas de IA no vulnera los derechos humanos, las libertades, la dignidad o la privacidad, deben establecerse marcos de gobernanza, incluidas normas y organizaciones reguladoras. Estos procesos también deben proporcionar trazabilidad. Esto contribuirá a aumentar la confianza del público en la IA [1].

Es necesario encontrar un mecanismo para convertir las consideraciones políticas y tecnológicas existentes y futuras, en obligaciones legales apegadas a derecho. Un procedimiento de este tipo debe tener en cuenta diversas normas culturales, así como diversos marcos jurídicos y normativos [5].

2. Bienestar

> Los creadores de sistemas autónomos e inteligentes adoptarán el aumento del bienestar humano como principal criterio de éxito para el desarrollo *[1, p. 21].*

Las mejores y más utilizadas mediciones de bienestar deberían aplicarse como referencia para los sistemas de IA, con el fin de garantizar que sea la prioridad como resultado en todos los diseños [6]. La declaración sugiere que, a la hora de diseñar e implantar la IA y los sistemas autónomos, es importante dar prioridad al bienestar humano como objetivo último. Para lograrlo, deben aplicarse indicadores ampliamente aceptados como referencia para evaluar el impacto de estos sistemas. Si bien esto no resulta tan sencillo en muchos casos, se puede crear las métricas adecuadas que contabilicen, de alguna manera, los impactos positivos y negativos que la tecnología pueda tener en los usuarios. Si se deja al azar, los aspectos perjudiciales podrían ser detectados cuando ya se es demasiado tarde.

3. Control de los datos

> Los creadores de sistemas autónomos e inteligentes deberán dotar a las personas de la capacidad de acceder a sus datos y compartirlos de forma segura, para mantener la facultad de las personas de tener control sobre su identidad *[1, p.23].*

Para IEEE los gobiernos y otras organizaciones deberían emprender esfuerzos para investigar, probar y aplicar tecnologías y procedimientos que permitan a los usuarios tener control sobre sus datos personales, en concreto permitiéndoles decidir caso por caso quién puede acceder a sus datos y procesarlos y con qué fines específicos [7]. Es crucial explorar si las estrategias de tutela existentes para niños y personas con capacidad de decisión disminuida se ajustan a esta recomendación o si alguien más debiese asumir la responsabilidad [8].

El ser humano debe controlar a los sistemas de IA sobre el modo en que toman decisiones, aprenden y el manejo que hagan de la información personal ante terceros. Por ejemplo, en todo momento permitir a los usuarios eliminar sus datos y su cuenta por completo.

4. Eficacia

> Los creadores y operadores deberán aportar pruebas de la eficacia y adecuación a los fines de los sistemas autónomos e inteligentes *[1, p. 25].*

Los responsables del diseño y la implantación de los sistemas de IA deben demostrar mediante pruebas tangibles que los sistemas son capaces de alcanzar los objetivos previstos y son adecuados para el fin perseguido. De acuerdo con la IEEE, la creación de sistemas de IA debe tener como objetivo la identificación de métricas o puntos de referencia que sirvan como indicadores fiables del éxito del sistema en la consecución de sus objetivos, el cumplimiento de las normas y el funcionamiento dentro de las tolerancias de riesgo [9]. Los diseñadores de sistemas de IA deben asegurarse de que todas las partes interesadas, como usuarios, certificadores de seguridad y reguladores del sistema, puedan acceder fácilmente a los resultados cuando se apliquen las métricas establecidas [10].

Considérense los sistemas de IA para la conducción autónoma, los desarrolladores y operadores deben aportar pruebas que demuestren que el software es eficaz para percibir con precisión el entorno, tomar decisiones oportunas y garantizar la seguridad de pasajeros y peatones. Estas pruebas deben incluir resultados de pretests rigurosos, estudios de simulación, datos de rendimiento en el mundo real y el cumplimiento de las normas de seguridad pertinentes antes de salir al mercado. Si bien hay lugares en que las normativas gubernamentales son más laxas, se debe tener altos estándares éticos en cada proyecto.

5. Transparencia

> La base de una decisión concreta de un sistema autónomo e inteligente debe poder descubrirse siempre *[1, p. 27].*

La IEEE considera que, al garantizar la detectabilidad, los desarrolladores y operadores de los sistemas de IA promueven la responsabilidad, la equidad y la capacidad de identificar y rectificar cualquier sesgo, error o consecuencia no deseada. La transparencia permite evaluar el proceso de toma de decisiones, lo que es crucial para la creación de confianza, la auditoría y la trazabilidad general.

Por ejemplo, los usuarios de robots domésticos o asistenciales deberían disponer de un botón que, al oprimirlo, el robot explique la acción que acaba de realizar [11]. También deben contar con un almacenamiento seguro de los datos de los sensores y del estado interno, similar al de una grabadora de datos de vuelo [12]. El Internet de las cosas ha levantado suspicacias precisamente por la opacidad en algunos casos, por ejemplo, en que se recojan y compartan datos personales sin el consentimiento explícito del usuario.

6. Responsabilidad

> Los sistemas autónomos e inteligentes deben crearse y gestionarse de forma que ofrezcan una justificación inequívoca de las decisiones adoptadas *[1, p. 29].*

Los sistemas autónomos e inteligentes pueden emprender acciones y tomar decisiones sin supervisión humana directa. Esto incluye los coches que se conducen solos, la robótica, los sistemas comerciales automatizados, etcétera. Estos sistemas deben crearse y gestionarse de forma responsable. Su diseño, desarrollo y despliegue deben seguir principios éticos capaces de justificar y explicar sus decisiones y acciones [5]. Las explicaciones deben ser claras e inequívocas, no vagas, contradictorias ni susceptibles de múltiples interpretaciones. No debe haber dudas sobre la lógica del comportamiento del sistema.

Esta explicabilidad es importante para mantener la responsabilidad, depurar el sistema, evitar sesgos involuntarios, generar confianza entre los usuarios y mucho más. Si un sistema no puede explicarse con claridad, resulta difícil evaluar si funciona de forma segura, ética o según lo previsto [6].

Esto es especialmente importante porque estas tecnologías son relativamente nuevas y aún no se conocen las repercusiones a futuro [13].

7. Conciencia del mal uso

> Los creadores deberán protegerse contra todos los usos indebidos y riesgos potenciales de sus sistemas autónomos e inteligentes en funcionamiento *[1, p. 31].*

Los diseñadores de sistemas de IA deben conocer las técnicas habituales de abuso y tratar de no hacerlas accesibles en sus diseños [14]. Ofrecer instrucción ética y hacer conciencia de los posibles riesgos del uso indebido de la tecnología de IA, y sensibilizar al público sobre las repercusiones. Por ejemplo, la creación de imágenes ficticias puede ser muy cómico para algunos, pero puede traer daños graves a la sociedad, a las personas involucradas y a la economía, aunque, en realidad, todavía no se pueda saber que alcances habrá de tener este tipo de fenómenos [7].

8. Competencia

> Los creadores harán las especificaciones y los operadores deberán respetar los conocimientos y destreza necesarios para un funcionamiento seguro y eficaz *[1, p. 32].*

Los tipos y grados de conocimiento necesarios para comprender y utilizar cualquier aplicación de IA deben ser especificados por sus creadores. Deben identificar los conocimientos requeridos, tanto para el sistema en su conjunto como para cada componente individual [15]. Los operadores de IA deben establecer políticas escritas que especifiquen cómo debe utilizarse. Estas directrices deben cubrir los usos prácticos de los sistemas, los requisitos previos para su uso eficaz, quién es competente para manejarlos, qué formación es necesaria para los operadores, cómo evaluar rendimiento del sistema y qué resultados deben esperarse. Las políticas también deben especificar las situaciones en las que puede ser necesario que el operador anule o detenga al sistema mismo [16].

Las recomendaciones anteriores están encaminadas a crear un estándar y conforman la primera serie de normas para una directriz tipo ISO sobre ética en la IA denominada serie IEEE P7000™ de proyectos de estandarización.

Norma IEEE 7000™ 2021. Preocupaciones éticas durante el diseño de sistemas

Más de 150 expertos trabajaron en la "Norma 7000™-2021 Proceso del modelo estándar IEEE para abordar las preocupaciones éticas durante el diseño del sistema" (7000™-2021IEEE Standard Model Process for Addressing Ethical Concerns during System Design) [17] y fue el resultado de debates en línea que tuvieron lugar a lo largo de cinco años y en los que participaron representantes de Europa, Oriente Medio, Estados Unidos, Australia y América Latina, así como especialistas de distintas disciplinas entre las que destacan ingenieros en sistemas, filósofos, consultores y abogados, entre otras áreas.

El modelo de proceso estándar del IEEE tiene como objetivo combinar temas éticos con la práctica para reducir riesgos e impulsar la innovación de la ingeniería de sistemas dentro de un enfoque compartido.

Para la IEEE ignorar los valores del usuario es un peligro en el diseño de ingeniería. Muchos bienes y servicios funcionan con sistemas de IA, que son algoritmos que operan "por debajo" del sistema y que tienen un impacto significativo en los datos, las identidades y los valores de los usuarios [18]. A pesar de los mejores esfuerzos de un fabricante, un proceso de diseño enfatizará las convicciones de sus creadores y su solvencia ética no siempre estará garantizada. En la era de los algoritmos, la innovación responsable requiere un enfoque basado en principios que vayan más allá de la ingeniería de sistemas convencional [19].

El estándar IEEE 7000™-2021 ofrece precisamente estos valores a las empresas a través de un método práctico para superar los problemas asociados a su transformación digital. La metodología ofrece una perspectiva más amplia para tener en cuenta los posibles daños causados por el diseño de productos o sistemas que no estén bien calibrados [20].

En el contexto del aprendizaje automático, se dice que un algoritmo está mal calibrado si sus predicciones o estimaciones son sistemáticamente sesgadas o inexactas. Más concretamente, un algoritmo mal calibrado puede producir estimaciones o predicciones que son, en promedio, demasiado altas o bajas en comparación con los valores reales [21]. Por ejemplo, considérese un escenario en el que un algoritmo está entrenado para evaluar la probabilidad de un resultado binario, como si un paciente tiene o no una enfermedad específica. Si el algoritmo sobrestima

sistemáticamente esta posibilidad, puede indicar erróneamente que está enfermo cuando, en realidad, no lo está [22].

Conclusiones parciales

La iniciativa de IEEE proporciona directrices éticas que se aplican a todo tipo de sistemas autónomos e inteligentes, incluidos los robots mecánicos, así como los robots algorítmicos, autos de conducción autónoma, sistemas de software, sistemas de diagnóstico médico, asistentes personales inteligentes y bots algorítmicos de chat, en diferentes entornos, tanto reales como virtuales, contextuales y de realidad mixta en donde la IA esté presente.

El texto de la Iniciativa y el Estándar 7000™-2021 pretenden ofrecer un marco para las consideraciones éticas específicas en el diseño, desarrollo, despliegue y uso de los sistemas de IA, con el objetivo de garantizar que estas tecnologías beneficien a la humanidad.

Algunas críticas al documento del IEEE es que, nuevamente como en otras iniciativas, los criterios son demasiado amplios y carecen de orientaciones específicas sobre cómo aplicar en la práctica los principios que esboza. Esto puede dificultar que organizaciones y gobiernos traduzcan los principios en acciones y políticas concretas. Sin embargo, también debe decirse que la IEEE es la organización que ha sido más incisiva al crear las normas correspondientes al desarrollo tecnológico, y prueba de ello es la “Norma 7000TM-2021 Proceso del modelo estándar IEEE para abordar las preocupaciones éticas durante el diseño del sistema” que no se debe subestimar como un esfuerzo puntual de implementación de la ética en la IA.

Por supuesto, también se ha expresado la preocupación de que los documentos del IEEE puedan estar influidos por los intereses de las partes interesadas de la industria, como las empresas tecnológicas y no representen plenamente las perspectivas e intereses de todas las partes, en particular las de las comunidades digitalmente marginadas.

A pesar de estas críticas, el Estándar 7000™-2021 y la Iniciativa son influyentes en el campo de la ética y la gobernanza de la IA. La IEEE desempeña un papel sustancial en la configuración del debate mundial en torno al desarrollo y despliegue responsables, no solo de la IA, sino del progreso científico de la humanidad.

Referencias

[23] IEEE. “Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition.” IEEE.org. Acceso abr. 2023. [En línea]. Disponible: https://bsu.buap.mx/b3f

[24] IEEE. “History of IEEE.” IEEE.org. Acceso abr. 2023. [En línea]. Disponible: https://www.ieee.org/about/ieee-history.html.

[25] IEEE. “Membership.” IEEE.org. Acceso abr. 2023. [En línea]. Disponible: https://www.ieee.org/membership/index.html.

[26] R. Chatila y J.C. Havens, “The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems,” en Robotics and Well-Being, M.I. Aldinhas Ferreira et al., Eds., Springer International Publishing, 2019, pp. 11-16, doi: 10.1007/978-3-030-12524-0_2

[27] T. Tzimas, “Legal and Ethical Challenges of Artificial Intelligence from an International Law Perspective.” Cham: Springer, 2021.

[28] M. Dubber, F. Pasquale, y S. Das, Eds. *The Oxford Handbook of Ethics of AI*. Oxford, UK: Oxford University Press, 2020.

[29] J. Antoniou y O. Tringides, “Personal Data, Cloud Platforms, Privacy and Quality of Experience,” en *Effects of Data Overload on User Quality of Experience*, J. Antoniou and O. Tringides, Eds., Cham: Springer International Publishing, 2023, pp. 37-54, doi: 10.1007/978-3-031-06870-6_3.

[30] S. Milano, M. Taddeo, y L. Floridi, “Recommender systems and their ethical challenges,” *AI & Soc.*, vol. 35, no. 4, pp. 957-967, 2020, doi: 10.1007/s00146-020-00950-y.

[31] T. Winkle, “Product Development within Artificial Intelligence, Ethics and Legal Risk.” Alemania: Springer Vieweg, 2022.

[32] L. Floridi, “Ethics, Governance, and Policies in Artificial Intelligence.” Cham: Springer, 2021.

[33] M. Coeckelbergh, “Robot ethics.” MA, USA: MIT Press, 2022.

[34] T. Hauer, “Incompleteness of moral choice and evolution towards fully autonomous AI,” Humanit. and soc. sciences commun., vol. 9, no. 1, p. 38, 2022. Disponible: https://bsu.buap.mx/ciT

[35] J. Bryson, “The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation,” en *The Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, y S. Das, Eds., Oxford University Press, 2020, pp. 2-25.

[36] K. Kumari y J. P. Singh, “AI ML NIT Patna at HASOC 2019: Deep Learning Approach for Identification of Abusive Content,” en *Proceedings of Forum for Information Retrieval Evaluation*, vol. 2517, pp. 328-335, 2019. Disponible: https://ceur-ws.org/Vol-2517/T3-20.pdf

[37] M. Groß, “Yes, AI Can: The Artificial Intelligence Gold Rush Between Optimistic HR Software Providers, Skeptical HR Managers, and Corporate Ethical Virtues,” en *AI for the Good: Artificial Intelligence and Ethics*, S.H. Vieweg, Ed., Cham: Springer International Publishing, 2021, pp. 191-225.

[38] B. Zhang et al., “Ethics and Governance of Artificial Intelligence: A Survey of Machine Learning Researchers,” en *IJCAI International Joint Conference on Artificial Intelligence*, 2022.

[39] IEEE, “7000-2021 - IEEE Standard Model Process for Addressing Ethical Concerns during System Design,” IEEE.org. Acceso abr. 2023. [En línea]. Disponible: https://ieeexplore.ieee.org/document/9536679.

[40] L. Floridi, “Establishing the Rules for Building Trustworthy AI,” en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Cham: Springer, 2021, pp. 41-45.

[41] L. Floridi y J. Cowls, “A Unified Framework of Five Principles for AI in Society,” en *Ethics, Governance, and Policies in Artificial Intelligence*, L. Floridi, Ed., Cham: Springer International Publishing, 2021, pp. 5-17.

[42] B.C. Stahl, “AI Ecosystems for Human Flourishing: The Recommendations,” en *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*, B.C. Stahl, Ed., Springer International Publishing, 2021, pp. 91-115.

[43] S. Russell y P. Norvig, “Philosophy, ethics, and safety of AI,” en *Artificial Intelligence: A Modern Approach*, Londres: Pearson, 2022, pp. 1032-1062.

[44] G.Z. Yang, et al., “The grand challenges of Science Robotics,” Sci Robot., vol. 3, no. 14, p. eaar7650, ene. 2018, doi: 10.1126/scirobotics.aar7650. PMID: 33141701.