



CONSEJO ASESOR INTERNACIONAL DE INTELIGENCIA ARTIFICIAL

Octubre 2026

DECLARACIÓN DE MADRID SOBRE UNA IA RESPONSABLE

Hacia una IA democrática, social, innovadora,
inclusiva, sostenible y responsable.



MIEMBROS DEL CONSEJO

Manuel Castells

Presidente. Catedrático de Tecnología de la Comunicación y Sociedad en la Universidad del Sur de California (USC) y catedrático emérito de la Universidad de California, Berkeley.

Luciano Floridi

Profesor John K. Castle de Práctica en Ciencias Cognitivas y director fundador del Digital Ethics Center de la Universidad de Yale; profesor de Sociología de la Cultura y de la Comunicación en la Universidad de Bolonia; experto en ética de la IA.

Paul S. Adler

Profesor Titular Emérito de la Cátedra Harold Quinton de Política Empresarial y Profesor Emérito de Gestión y Organización en la Marshall School of Business de la Universidad del Sur de California (USC).

Jeroen van den Hoven

Catedrático de Ética y Tecnología en la Universidad Tecnológica de Delft y director científico del Instituto Delft Design for Values.

Francesca Bria

Profesora Honoraria del Institute for Innovation and Public Purpose de UCL, Presidenta de ART-ER (la agencia de innovación de Emilia-Romaña) y miembro del Consejo del European Innovation Council; experta en soberanía digital.

Carissa Véliz

Profesora asociada en el Instituto de Ética en IA de la Universidad de Oxford y autora de *Privacidad es poder*.

Manifiesto de Madrid por una IA responsable

Hacia una IA democrática, social, innovadora, inclusiva, sostenible y responsable

Vint Cerf

Pionero de Internet y galardonado con importantes reconocimientos, entre ellos la Medalla Nacional de Tecnología de Estados Unidos y el Premio Turing.

Erika Staël von Holstein

Directora ejecutiva del think tank independiente Re-Imagine Europa; experta en inteligencia artificial y polarización política.

Kate Crawford

Profesora en la USC Annenberg; titular de la Cátedra Fundacional de IA y Justicia de la École Normale Supérieure; destacada experta en las implicaciones sociales, políticas y medioambientales de la inteligencia artificial.

Niklas Lundblad. Experto en políticas tecnológicas y tecnología responsable, y antiguo directivo global de Google, con más de veinte años de experiencia en políticas tecnológicas e investigación.

Jerome A. Feldman. Profesor emérito de Ingeniería Eléctrica y Ciencias de la Computación en la Universidad de California, Berkeley; figura destacada en computación y ciencia cognitiva.

CONTEXTO

El Consejo Asesor Internacional de Inteligencia Artificial fue creado en junio de 2024 con el propósito de ayudar al Gobierno de España a afrontar uno de los grandes desafíos de nuestro tiempo: comprender y abordar los retos y oportunidades que plantea la inteligencia artificial, orientando su desarrollo hacia el progreso compartido, la protección de los derechos y el fortalecimiento de la democracia.

Presidido por Manuel Castells e integrado por expertos de reconocido prestigio internacional, el Consejo reúne una amplia diversidad de disciplinas y perspectivas para proporcionar asesoramiento estratégico independiente sobre las implicaciones a largo plazo de la inteligencia artificial.

El Manifiesto de Madrid es el resultado del trabajo desarrollado por el Consejo desde su constitución en 2024 y refleja las reflexiones, deliberaciones y recomendaciones elaboradas a lo largo de sus sucesivas reuniones. El texto reúne las posiciones consensuadas por sus miembros y expresa una convicción compartida: la inteligencia artificial alcanzará su verdadero potencial cuando las sociedades sean capaces de integrarla de manera efectiva en beneficio de las personas, fortaleciendo la cohesión social y ampliando las oportunidades de progreso. El Manifiesto aspira así a convertir el conocimiento acumulado por el Consejo en una referencia común para una inteligencia artificial democrática, innovadora, inclusiva, sostenible y responsable.

DESCARGO DE RESPONSABILIDAD

El presente Manifiesto refleja el trabajo colectivo y el amplio consenso de los miembros del Consejo Asesor Internacional de Inteligencia Artificial. Los miembros participan a título personal y no como representantes de ningún Gobierno, institución u organización. Aunque el Manifiesto expresa una visión compartida y unos principios comunes acordados por el Consejo, no debe entenderse que todos los miembros respalden necesariamente en su totalidad todas y cada una de las afirmaciones que contiene.

Nada de lo dispuesto en este Manifiesto podrá interpretarse como expresión de la posición oficial de ningún Gobierno, institución u organización con la que pueda estar vinculado alguno de los miembros del Consejo. Asimismo, el presente Manifiesto no constituye ni representa una posición oficial del Gobierno de España.

ÍNDICE

Preámbulo

I. Una inteligencia artificial humanocéntrica al servicio de las personas y del bien común

- IA humanocéntrica
- Respeto y protección de derechos humanos y libertades civiles
- Defensa y fortalecimiento de la autonomía humana
- Igualdad, inclusión y no discriminación
- Privacidad y minimización de datos
- Protección reforzada de menores

II. Una inteligencia artificial soberana y democrática

- Protección y fortalecimiento del espacio público
- Soberanía digital
- Autonomía cognitiva
- Resiliencia
- Capacidades institucionales, públicas y sociales
- Infraestructuras digitales al servicio del interés público

III. Una inteligencia artificial inclusiva, innovadora y sostenible

- Finalidad legítima
- Proporcionalidad tecnológica
- Un concepto amplio de progreso
- Educación y formación
- Mejora del trabajo
- Diversidad social, lingüística y cultural
- Sostenibilidad intergeneracional, material y territorial

IV. Una inteligencia artificial responsable, transparente y fiable

- Identidad y transparencia
- Confianza y fiabilidad
- Trazabilidad, auditoría y reparación
- Evaluación durante todo el ciclo de vida
- Control humano efectivo
- Justicia y responsabilidad

Manifiesto de Madrid por una IA responsable

Hacia una IA democrática, social, innovadora, inclusiva, sostenible y responsable

PREÁMBULO

La inteligencia artificial (IA) proporciona posibilidades reales de progreso. Si se diseña y se implementa correctamente, puede acelerar el descubrimiento científico y el diagnóstico médico, ampliar el acceso al conocimiento y a la educación, mejorar la calidad de los servicios públicos y permitir a las personas con discapacidad participar en condiciones más justas y equitativas. Esas oportunidades son precisamente la razón por la que hay que desarrollarla bien.

La IA es, además, un recurso geopolítico y una infraestructura transversal, social, económica, informativa y política, que transforma el conjunto de los sectores y de la actividad. Por ello hay que impulsar una IA que sea compatible con los principios y valores de las sociedades democráticas y que contribuya a su fortalecimiento.

La capacidad de decisión sobre la IA se concentra cada vez más en manos de un reducido número de grandes actores empresariales y tecnológicos, que ejercen de facto un poder sobre las condiciones de la vida cotidiana, monopolizando los conocimientos, los datos y las herramientas digitales. Empresas y plataformas determinan las condiciones de acceso, las reglas de visibilidad, las formas de interacción e incluso parte de las oportunidades económicas. En la era digital, el colonialismo adopta nuevas formas. Ya no domina únicamente los cuerpos, sino que se apropia de los datos, transformando las vidas personales en información explotable. La libertad ya no es solo un asunto interno de la esfera individual de la persona, sino también una cuestión de interés público. Por ello, son necesarias reglas claras, transparencia, mecanismos efectivos de recurso y límites proporcionados al uso de tecnologías intrusivas, para asegurar que la tecnología permanezca al servicio de las personas.

Esta situación plantea desafíos fundamentales para la gobernanza democrática, la equidad económica y la soberanía tecnológica. La autorregulación resulta insuficiente para garantizar que el desarrollo de la IA responda al interés general. Por ello, las sociedades democráticas tienen un papel legítimo en la orientación, regulación, gobernanza y supervisión de la IA, a fin de asegurar que sus beneficios, oportunidades

y capacidades se distribuyan ampliamente y permanezcan alineados con los derechos humanos y los valores democráticos.

La transformación digital de las últimas décadas ha enseñado valiosas lecciones, sobre todo, acerca de cómo el poder y el valor tienden a concentrarse cuando un pequeño número de actores controla las infraestructuras de las que dependen los demás. La IA ofrece la oportunidad de aplicar estas lecciones y de avanzar de modo diferente.

La IA puede convertirse en un medio para concentrar conocimiento y capacidad de decisión en un pequeño número de agentes o en un medio para que muchas personas se apoyen en esos conocimientos y desarrollen soluciones adaptadas a sus propias necesidades y contextos. El camino que tomemos depende de unos fundamentos que se están estableciendo ahora: cómo se accede a los datos, cómo se reconoce y recompensa a los creadores y a otros contribuyentes, cómo se confía en la información y cómo los sistemas se conectan entre sí. Estos fundamentos se configuran mejor de manera abierta. Sobre reglas claras e infraestructuras de confianza accesibles para todos, la IA puede convertirse en una fuente de creatividad y oportunidades para millones de personas: para los cientos de miles de pequeñas y medianas empresas, empresas emergentes, creadores, investigadores y comunidades de toda Europa y más allá que pueden utilizarla para resolver los problemas concretos en sus sectores de actividad y sus territorios.

El objetivo que compartimos es orientar, regular y gobernar la IA para que el conocimiento, el poder y los beneficios que genera fortalezcan la autonomía personal, contribuyan al bienestar y a la cohesión económica y social y se mantengan claramente alineados con los valores democráticos y el interés público general. Más allá de las capacidades de los sistemas, el éxito depende de la opción por una IA humanocéntrica, democrática, innovadora, inclusiva, sostenible, responsable y orientada al interés público. Esa opción está a nuestro alcance.

A efectos del presente Manifiesto, toda referencia a la IA se entenderá también referida a quienes diseñan, desarrollan, despliegan, implementan, operan y gobiernan sistemas de IA.

SECCIÓN I.

UNA INTELIGENCIA ARTIFICIAL HUMANOCÉNTRICA AL SERVICIO DE LAS PERSONAS Y DEL BIEN COMÚN

01 IA Humanocéntrica

La IA debe estar al servicio de las personas, la sociedad y el bien común. La tecnología debe complementar la inteligencia, la creatividad, la experiencia profesional y el bienestar de las personas.

02 Respeto y protección de los derechos humanos y las libertades civiles

Los derechos humanos y las libertades civiles deben incorporarse desde el diseño y durante todo el desarrollo, despliegue e implementación de los sistemas de IA.

Esos derechos y libertades, junto a los principios éticos, deben traducirse en requisitos técnicos y organizativos verificables.

03 Defensa y fortalecimiento de la autonomía humana

La IA debe respetar y fomentar la autonomía humana como capacidad y libertad para pensar, actuar, decidir y controlar la propia vida.

Las predicciones y recomendaciones generadas por sistemas de IA pueden servir de apoyo a decisiones que afecten de forma significativa a las personas, pero deben mantenerse claramente diferenciadas de estas, de modo que la responsabilidad de cada decisión recaiga en una persona o institución identificable, capaz de explicarla y de responder por ella. Debe preservarse una separación clara y efectiva entre predicción, recomendación y decisión.

04 Igualdad, inclusión y no discriminación

La IA debe diseñarse, desarrollarse, implementarse y evaluarse con el propósito de contribuir a la igualdad, favorecer la inclusión y garantizar la no discriminación.

La IA puede ser una herramienta especialmente valiosa para reducir las desigualdades derivadas de la falta de recursos, competencias, conectividad o accesibilidad o de otra naturaleza.

05 Privacidad y minimización de datos

La privacidad es un derecho individual y una condición de libertad y democracia. La vigilancia masiva entra en conflicto con los derechos humanos y las libertades civiles. Proteger la privacidad en el uso de la IA es una forma de favorecer un reparto justo y equilibrado del poder.

La minimización de datos es un principio básico. Todo tratamiento de datos debe responder a un objetivo legítimo y solo deben utilizarse aquellos datos estrictamente necesarios para alcanzarlo.

06 Protección reforzada de menores

La protección de los menores no es una limitación a la innovación, sino la condición que la legitima. Los niños, niñas y adolescentes son titulares de derechos y su interés superior debe guiar el diseño, desarrollo y despliegue de la IA.

Los asistentes, los acompañantes de IA y los chatbots deberían diseñarse para apoyar el bienestar, la autonomía y el desarrollo saludable de las personas menores. Los sistemas de IA susceptibles de llegar a personas menores deberían someterse a evaluaciones independientes antes de su despliegue, con el fin de garantizar elevados estándares de seguridad, protección y rendición de cuentas.

Toda tecnología transforma el mundo que heredarán las generaciones futuras. Nuestro futuro y el de nuestros hijos deben situarse por delante de los beneficios de la industria de la IA. La responsabilidad de esta generación es asegurar que la IA amplíe sus oportunidades mediante la seguridad, un diseño adecuado para cada edad, la alfabetización digital y unos resultados educativos positivos.

SECCIÓN II.

UNA INTELIGENCIA ARTIFICIAL SOBERANA Y DEMOCRÁTICA

07 Protección y fortalecimiento del espacio público

La IA debe fortalecer la democracia y el espacio público, contribuyendo a crear condiciones para que las personas formen libremente sus opiniones y participen en la vida pública.

La IA puede utilizarse para mejorar la deliberación democrática: resumir posiciones enfrentadas con fidelidad, detectar coincidencias o facilitar consultas públicas multilingües. La IA puede contribuir positivamente a la democracia. Para que la IA pueda desempeñar este papel, los canales a través de los cuales las personas obtienen cada vez más información, incluidos los sistemas de IA conversacional, deben seguir siendo plurales y estar sujetos a mecanismos de responsabilidad y rendición de cuentas.

08 Soberanía digital

La soberanía digital combina tres dimensiones inseparables: democracia, seguridad y competitividad. Garantiza la autodeterminación informativa de la ciudadanía, mantiene las infraestructuras digitales críticas bajo supervisión democrática y al amparo del Derecho europeo y permite desarrollar y gobernar capacidades tecnológicas clave para que los beneficios de la IA se compartan ampliamente en toda la sociedad.

La soberanía digital se construye mediante la apertura y la colaboración. Significa garantizar que las personas y las instituciones mantengan un control efectivo sobre los datos, sistemas y servicios de los que dependen y que la IA refleje la diversidad lingüística, cultural y territorial de las sociedades. Los sistemas que incorporan una variedad de lenguas, contextos y conocimientos prestan un mejor servicio a la ciudadanía. Los estándares abiertos y la interoperabilidad permiten que esa diversidad se convierta en una fortaleza compartida.

La apertura, la interoperabilidad y la cooperación internacional son plenamente compatibles con la autonomía estratégica cuando preservan la capacidad de las sociedades democráticas para adoptar sus propias decisiones y definir su futuro tecnológico.

La soberanía digital es una condición necesaria para la democracia. Es la capacidad de las sociedades democráticas para gobernar las infraestructuras digitales críticas de acuerdo con sus normas, valores y prioridades colectivas. En la era digital, la soberanía no significa aislamiento, sino autonomía, resiliencia y democracia.

La infraestructura de IA constituye una base estratégica para la prosperidad económica, la resiliencia democrática y la cohesión social. Las sociedades democráticas deben mantener capacidades públicas sólidas en todas las capas de la cadena de valor de la IA para garantizar que las infraestructuras tecnológicas críticas permanezcan bajo supervisión democrática y al servicio del interés general. La capacidad pública constituye una piedra angular de la soberanía democrática en la era digital.

09 Autonomía cognitiva

La autonomía cognitiva es la capacidad de las personas para formar sus propias convicciones y adoptar sus propias decisiones, preservando su libertad de elección frente a interferencias indebidas derivadas de la tecnología o de los sistemas digitales.

Fundamentada en la libertad de pensamiento y en el derecho a la integridad mental, la autonomía cognitiva debe fortalecerse mediante sistemas de IA que refuercen la autonomía humana, contribuyan al bienestar y faciliten una toma de decisiones informada. Los sistemas de IA también pueden amenazar la autonomía cognitiva, especialmente cuando buscan manipular emociones y creencias, simular amistad o explotar vulnerabilidades psicológicas. Quienes desarrollan y despliegan sistemas de IA no deben diseñarlos ni operarlos de forma que comprometan la autonomía cognitiva o maximicen la interacción a costa del bienestar de las personas y de su libertad de pensamiento.

Cuando intervengan neurotecnologías, estos principios deberán complementarse con las salvaguardias comúnmente denominadas neuroderechos.

10 Resiliencia

La resiliencia consiste en proteger los sistemas frente a fallos o ataques y en preservar la capacidad de las personas, las instituciones y las sociedades para seguir funcionando, adaptarse y prosperar en un entorno de cambio tecnológico acelerado.

La tecnología debe contribuir a reforzar la resiliencia de la sociedad, nunca a debilitarla.

La resiliencia convierte la innovación en progreso sostenible y garantiza que los beneficios puedan mantenerse frente a crisis y transformaciones futuras.

11 Capacidades institucionales, públicas y sociales

La capacidad pública implica fomentar que las instituciones cuenten con el conocimiento suficiente para comprender, evaluar, adquirir, supervisar y, en su caso, desarrollar sistemas de IA. Es una inversión en autonomía democrática porque permite decidir cómo utilizar la tecnología.

12 Infraestructuras digitales al servicio del interés público

Las infraestructuras digitales públicas constituyen el fundamento estructural de una IA inclusiva, resiliente y al servicio del interés general. Invertir en ellas significa invertir en la capacidad colectiva de innovar, competir y decidir autónomamente en la economía del conocimiento.

La contratación pública debe desempeñar un papel estratégico en la consecución de este objetivo, promoviendo la apertura, la interoperabilidad, la seguridad y la sostenibilidad y favoreciendo el desarrollo de alternativas fiables en las capas críticas de la infraestructura digital.

La acción pública no sustituye a la innovación privada; crea las condiciones para que pueda prosperar en un entorno abierto y competitivo.

Los ecosistemas de IA abiertos e interoperables, que priorizan los open weight models, las herramientas y los estándares abiertos, desempeñan un papel esencial para impulsar la innovación, el progreso científico, la transparencia, la competencia y el acceso democrático a la tecnología. Las políticas públicas deben apoyar su desarrollo y adopción, garantizando al mismo tiempo salvaguardias adecuadas en materia de seguridad, protección y derechos fundamentales.

Más allá de la infraestructura física, la economía de la IA necesita unos cimientos comunes que todavía están en gran medida por construir: estándares interoperables; una identidad fiable para las personas, las organizaciones y los agentes de IA que actúan en su nombre; señales de procedencia y confianza para la información; mecanismos que reconozcan y remuneren a los creadores y demás contribuyentes; sistemas seguros y resilientes; y un acceso equitativo a la capacidad de cómputo, los datos y los modelos. Las instituciones públicas deberían contribuir a construir estos cimientos junto con investigadores, sociedad civil, trabajadores, creadores y empresas de todos los tamaños, para que sigan siendo abiertos y accesibles para todos.

El control sobre los datos debe permanecer en manos de quienes los generan o los custodian. Siempre que sea posible, los sistemas de IA deberían acceder a los datos conforme a las condiciones establecidas por sus titulares, para una finalidad declarada y con un registro de uso, en lugar de extraerlos masivamente. Diferenciar la custodia de los datos de la propiedad de los modelos abre los mercados a un mayor número de desarrolladores, protege la privacidad y el trabajo creativo y favorece una competencia justa.

Dado que nadie tiene todavía todas las respuestas, la gobernanza de estos sistemas debe ser adaptativa: supervisando sus efectos, probando abiertamente nuevos enfoques, aprendiendo de la experiencia práctica y compartiendo los resultados entre países y comunidades.

SECCIÓN III.

UNA INTELIGENCIA ARTIFICIAL INCLUSIVA, INNOVADORA Y SOSTENIBLE

13 Finalidad legítima

El recurso a la IA debe responder a una necesidad legítima, concreta y claramente identificable, aportar un beneficio social verificable y contribuir de manera efectiva al bienestar de las personas, la protección de sus derechos o la consecución de objetivos de interés público.

La exigencia de una finalidad legítima no solo orienta los usos beneficiosos de la IA, sino que también establece límites claros a su despliegue. Los usos de la IA que sean intrínsecamente incompatibles con la dignidad humana, los derechos fundamentales y la paz y la seguridad internacionales no pueden considerarse legítimos.

En particular, las armas autónomas letales son moralmente inaceptables y ninguna máquina debería decidir jamás acabar con una vida humana. Debe existir un control humano efectivo, significativo y sujeto a rendición de cuentas sobre todos los sistemas de armas, así como una prohibición de las armas autónomas letales.

14 Proporcionalidad tecnológica

La utilización de sistemas de IA debe ser proporcional a la necesidad que se pretende atender, a los riesgos que genera y a los recursos que exige.

La proporcionalidad debe evaluarse atendiendo no solo a los beneficios esperados, sino también a sus efectos sobre los derechos, la autonomía, la sostenibilidad ambiental, el consumo de recursos y la distribución de costes y beneficios.

15 Un concepto amplio de progreso

La IA debe contribuir a una concepción amplia del progreso humano, entendida no solo como crecimiento económico o avance tecnológico, sino también como ampliación de capacidades, fortalecimiento de derechos, mejora del bienestar, cohesión social, sostenibilidad ambiental y profundización democrática.

La transformación digital debe favorecer una distribución inclusiva y equilibrada

del valor creado entre quienes participan en el ecosistema de datos, modelos, infraestructuras, trabajo, servicios y territorio. La negociación colectiva, una fiscalidad justa y la participación pública en las infraestructuras de IA constituyen mecanismos clave para alcanzar este objetivo.

El valor debe fluir hacia quienes contribuyen a crearlo, incluidos los creadores, investigadores, trabajadores, comunidades y empresas cuyos conocimientos y trabajo alimentan los sistemas de IA y no únicamente hacia quienes controlan el punto final de distribución.

La IA es mucho más que una carrera por el mejor modelo; es una carrera por construir la sociedad más capaz de aprovechar la IA y beneficiarse de ella. La competitividad de Europa en la era de la IA reside en sus millones de pequeñas y medianas empresas, empresas emergentes, investigadores y creadores, así como en unos mercados abiertos y disputables en los que todos ellos puedan desarrollar soluciones con IA. Unas normas claras, predecibles y proporcionadas, acompañadas de orientación práctica, entornos de pruebas regulatorios (sandboxes) y vías sencillas de cumplimiento, ofrecen a las empresas de todos los tamaños la confianza necesaria para experimentar, invertir y crecer.

16 Educación y formación

La educación y la formación permanente son fundamentales en la era de la IA. La integración social exige empoderar a las personas. La educación consiste en enseñar a utilizar nuevas herramientas y, también, en fortalecer la capacidad de las personas para pensar críticamente, distinguir entre información y manipulación, formular juicios propios y participar de manera informada en la vida social y democrática.

El éxito en la era de la IA dependerá menos de quién tenga la mejor tecnología y más de quién consiga integrar mejor esa tecnología en su sociedad.

17 Mejora del trabajo

La transformación tecnológica debe orientarse prioritariamente a complementar las capacidades de trabajadores y profesionales, mejorar la calidad del empleo y favorecer organizaciones más colaborativas y eficaces. Las ganancias de productividad derivadas de la IA deberían compartirse con los trabajadores mediante mejoras salariales y de las condiciones laborales, incluidas oportunidades para reducir el tiempo de trabajo sin pérdida salarial.

La transformación digital debe ir acompañada de medidas que garanticen la protección de los trabajadores, el acceso a la formación continua, el diálogo social, unas condiciones laborales justas, medidas para prevenir desplazamientos laborales evitables, apoyo a la recolocación profesional, seguridad de ingresos y

salvaguardias frente a la monitorización desproporcionada y otros impactos adversos derivados del uso de la IA. Estas protecciones deberían extenderse a los trabajadores de toda la cadena de valor de la IA, incluidos los trabajadores subcontratados, los etiquetadores de datos y los moderadores de contenido.

El resultado no viene determinado por la tecnología, sino por las decisiones organizativas que la acompañan: quién participa en su diseño, con qué objetivos se implanta y cómo se reparten sus beneficios. De esas decisiones depende que la automatización amplíe las capacidades profesionales.

Los trabajadores y sus representantes deberían tener derechos efectivos para participar en las decisiones relacionadas con el uso de sistemas algorítmicos en el ámbito laboral, entre otros medios mediante el diálogo social, la negociación colectiva y los procedimientos de información y consulta.

18 Diversidad social, lingüística y cultural

La IA debe contribuir a preservar y fortalecer la diversidad lingüística, cultural y territorial de las sociedades. Los sistemas de IA deben reconocer y representar adecuadamente las distintas lenguas, culturas, tradiciones de conocimiento y formas de expresión, ampliando la participación y evitando nuevas formas de exclusión.

19 Sostenibilidad

La IA debe contribuir a una visión amplia y duradera de la sostenibilidad. La IA debería generar prosperidad, conocimiento y bienestar para las generaciones presentes y futuras, impulsar un uso más eficiente e inteligente de los recursos materiales, fortalecer la resiliencia de nuestras infraestructuras y favorecer un desarrollo equilibrado de todos los territorios, respetando al mismo tiempo los recursos naturales y mejorando nuestra huella ecológica.

La sostenibilidad también exige que la huella física de la IA, desde la energía, el agua, el suelo y los materiales consumidos por los centros de datos y el hardware hasta los residuos que generan, sea medida, divulgada de forma transparente, mantenida dentro de los límites ambientales y sometida a una evaluación continua por parte de las comunidades y desde el punto de vista ambiental.

La sostenibilidad también tiene una dimensión financiera. Dado que las infraestructuras de IA se financian cada vez más mediante una combinación de deuda, crédito privado y otros instrumentos financieros, resulta esencial garantizar la transparencia en la distribución de riesgos y responsabilidades. Por ello, la inversión pública y las asociaciones para el desarrollo de infraestructuras de IA deberían ir acompañadas de una divulgación clara de quién asume en última instancia dichos riesgos.

SECCIÓN IV.

UNA INTELIGENCIA ARTIFICIAL RESPONSABLE, TRANSPARENTE Y FIABLE

20 Identidad y transparencia

La transparencia debe ser operativa y debe permitir comprender la finalidad del sistema, sus límites, riesgos, datos relevantes, criterios de funcionamiento, grado de intervención humana y los canales disponibles para presentar reclamaciones y solicitar reparación.

A medida que los sistemas de IA adquieren una capacidad cada vez mayor para actuar de forma autónoma, las sociedades digitales confiables necesitan infraestructuras de identidad interoperables que permitan identificar quién actúa, en nombre de quién lo hace y con qué autoridad. Los agentes de IA deben poder ser identificados de manera única y sus acciones deben ser atribuibles, mientras que la responsabilidad por dichas acciones debe recaer siempre en personas físicas o jurídicas claramente responsables y sujetas a rendición de cuentas.

Las sociedades democráticas deberían promover estándares de identidad abiertos, interoperables y respetuosos con la privacidad que refuercen la confianza, la seguridad y la rendición de cuentas en los ecosistemas digitales.

21. Confianza y fiabilidad

La confianza posibilita la adopción generalizada de la IA, ya que anima a las personas, las organizaciones y la sociedad a utilizarla cuando se demuestra que es segura, justa y comprensible. La confianza solo está justificada cuando deriva de una fiabilidad demostrada; por ello, el objetivo debe ser una IA digna de confianza. Cuando la IA es digna de confianza, las personas y las organizaciones pueden apoyarse en ella y sus beneficios pueden extenderse al conjunto de la economía.

La fiabilidad exige reconocer las limitaciones de la IA y garantizar una capacidad efectiva para corregir, remediar y retirar sistemas cuando sea necesario.

La fiabilidad debe evaluarse teniendo en cuenta los errores, las incertidumbres, la falta de contexto y las limitaciones del sistema, desde el diseño hasta su retirada definitiva.

22. Trazabilidad, auditoría y reparación

Los sistemas deben ser trazables en relación con datos, modelos, resultados, responsables, actualizaciones e incidentes.

La trazabilidad debe permitir detectar y corregir errores, identificar y reparar daños, reconstruir las decisiones relevantes, verificar el cumplimiento de los requisitos, establecer responsabilidades e implementar medidas correctivas.

Los sistemas deben contar con mecanismos claros para su suspensión, modificación, impugnación o retirada cuando sus resultados comprometan la seguridad o los derechos.

La responsabilidad por los daños causados por los sistemas de IA debe recaer en quienes los desarrollan y despliegan: cuando los sistemas de alto riesgo causen daños, la carga de la prueba debe recaer en los desarrolladores y responsables del despliegue. Las personas afectadas deben poder solicitar reparación colectivamente. Quienes informen desde dentro sobre riesgos o conductas indebidas deben estar protegidos.

23. Evaluación durante todo el ciclo de vida

La evaluación de los sistemas de IA debe acompañar todo su ciclo de vida, desde su diseño y desarrollo hasta su despliegue y utilización, garantizando que son seguros, fiables, útiles y coherentes con los objetivos y valores que justificaron su adopción.

Esta evaluación se beneficia de la cooperación entre quienes desarrollan los sistemas, quienes los evalúan de forma independiente, las autoridades públicas, los trabajadores, los usuarios y las comunidades afectadas.

Los sistemas de IA de frontera y los sistemas agénticos deben someterse a evaluaciones independientes antes de su despliegue, garantizando a los evaluadores acceso a los modelos, a la documentación de los datos de entrenamiento y a los registros de incidentes. Las autoridades públicas, y no únicamente los operadores, deben contar con la facultad de suspender o retirar cualquier sistema cuyo comportamiento escape al control humano efectivo.

24. Control humano real y efectivo

La supervisión humana debe ser real y efectiva y estar respaldada por los conocimientos, el contexto, la información, el tiempo, la autoridad y la capacidad para comprender, corregir, rechazar o detener el resultado. La creciente autonomía de la IA agéntica incrementa la necesidad de garantizar mecanismos de supervisión

humana. Estos requisitos se aplican también a las autoridades públicas, cuyo uso de la IA, especialmente en ámbitos como la seguridad, los servicios sociales y la justicia, debe estar sujeto a los más altos estándares de transparencia, posibilidad de impugnación y control humano efectivo.

El despliegue de agentes de IA debe apoyarse en infraestructuras interoperables y respetuosas con la privacidad que permitan acreditar la identidad, la representación y la responsabilidad de las personas u organizaciones en cuyo nombre actúan, garantizando así la transparencia y la rendición de cuentas.

25. Justicia y responsabilidad

La IA debe contribuir a una justicia más accesible, eficaz y transparente, fortaleciendo la responsabilidad en todas las decisiones que afectan a las personas. Su utilización en este ámbito debe basarse en el respeto al Estado de derecho, las garantías procesales y los hechos verificables.

Los sistemas de IA pueden prestar apoyo en funciones técnicas y administrativas, pero las decisiones judiciales y las decisiones que afecten significativamente a los derechos de las personas deben permanecer sujetas a una supervisión y responsabilidad humanas significativas.

Declaración final

Este Manifiesto expresa una convicción compartida: la IA debe estar al servicio de toda la humanidad. Apoyamos límites claros, mecanismos efectivos de rendición de cuentas y una distribución equitativa de los beneficios de la IA, preservando al mismo tiempo los derechos humanos, la autonomía y la dignidad de las personas.

Nuestro objetivo no es únicamente gobernar la IA de manera responsable, sino crear las condiciones para que las personas, las comunidades, los investigadores, los creadores y las empresas de todos los tamaños puedan decidir si, cómo y dónde desarrollan la IA, de forma que fortalezcan la democracia, amplíen las oportunidades y contribuyan a una prosperidad compartida. Cuando la IA es digna de confianza y sus fundamentos permanecen abiertos y accesibles, sus beneficios pueden extenderse al conjunto de la economía y de la sociedad.

