Global Coalition for Tech Justice

Propuesta para la Cumbre de Acción sobre Inteligencia Artificial

10 y 11 de febrero de 2025, Francia

Grupo de Trabajo sobre Inteligencia Artificial e Integridad de la Información
Coalición Global por la Justicia Tecnológica (GCTJ)
Noviembre 2024

Temas: Confianza en la IA, Gobernanza mundial de la IA e IA de interés público

Estas propuestas han sido elaboradas por el Grupo de Trabajo sobre IA de la Coalición Global por la Justicia Tecnológica.

La coalición reúne a 250 organizaciones miembros y expertos de 55 países. Su objetivo es garantizar que las grandes empresas tecnológicas cumplan su papel en la protección de la democracia y los derechos humanos en todo el mundo, especialmente en la mayoría global, donde las empresas han sido negligentes a la hora de hacer frente a los impactos de las plataformas y tecnologías.

Se listan los co-firmantes individuales, pero la propuesta permanecerá abierta a la adhesión hasta febrero de 2025.

Digital Action convoca y organiza la Coalición Global por la Justicia Tecnológica. Desde 2019, Digital Action moviliza una red global de aliados para exigir mejores estándares a los gobiernos y corporaciones responsables de nuestros entornos digitales.

Índice

I. Introducción

Desde el diseño hasta el despliegue, desde la formulación de políticas hasta la rendición de cuentas, existe una profunda desigualdad global en el corazón de la Inteligencia Artificial (IA), que está configurando cada vez más el futuro de la integridad de la información, la democracia y los derechos humanos.[1] Esta desigualdad representa un reto fundamental para la justicia global y la gobernanza democrática en la era digital. A medida que los sistemas de IA se integran más profundamente en instituciones e infraestructuras públicas críticas[2]—desde la sanidad y la educación hasta los sistemas judiciales y los servicios públicos—, este desequilibrio de poder amenaza aumentar las disparidades mundiales existentes y crear nuevas formas de dependencia tecnológica que debilitan la autonomía y la autodeterminación nacionales. La rápida aceleración del desarrollo de la IA, concentrada en unos pocos centros mundiales de poder,[3] corre el riesgo de asentar estas desigualdades en los cimientos de nuestro futuro digital compartido.

La IA se diseña principalmente en el Norte Global o en China y se implementa en otras regiones con una mínima consideración por los contextos o consecuencias locales. Los daños en la Mayoría Global quedan sistemáticamente sin abordar, mientras que sigue habiendo poca capacidad y experiencia para la elaboración de políticas de IA respetuosas con los derechos en estas regiones. Estos daños van desde el sesgo algorítmico[4] hasta el desplazamiento de los sistemas locales de toma de decisiones. Los efectos se reflejan de múltiples maneras: sistemas de IA que no reconocen los idiomas locales y los matices culturales, sistemas automatizados de toma de decisiones entrenados en datos occidentales que toman determinaciones inapropiadas en contextos del Sur Global, y sistemas de moderación de contenidos impulsados por IA que suprimen inadvertidamente el discurso político legítimo.[5] La falta de conocimientos y recursos locales para identificar y abordar estos daños, junto con la exclusión sistemática por parte de las empresas de los conocimientos y la experiencia del Sur Global en el desarrollo y despliegue de la IA, crea un ciclo de dependencia tecnológica y marginalización que se refuerza a sí mismo.

---

[1] United Nations, 'Urgent Action Needed over Artificial Intelligence Risks to Human Rights' (UN News, 17 September 2021) https://news.un.org/en/story/2021/09/1099972 consultado el 5 de mayo de 2024.

[2] Las Tecnologías de Utilidad General (TGF) "son tecnologías que, a lo largo de la historia, han cambiado toda la economía y, por tanto, tienen el potencial de aplicar cambios drásticos en la sociedad con un impacto en las estructuras económicas y sociales preexistentes". André Guidetti, Artificial Intelligence as General Purpose Technology: An Empirical and Applied Analysis of its Perception (Master's Thesis, Università della Valle d'Aosta - Université de la Vallée d'Aoste 2020), p.1 https://univda.unitesi.cineca.it/bitstream/20.500.14084/428/1/ETI_104_Guidetti_André.pdf consultado el 7 de octubre de 2024.

[3] Anu Bradford, 'The Race to Regulate Artificial Intelligence' (Foreign Affairs, 27 June 2023) https://www.foreignaffairs.com/united-states/race-regulate-artificial-intelligence-sam-altman-anu-bradford consultado el 25 de octubre del 2024.

[4] United Nations, 'Impact of New Technologies on the Promotion and Protection of Human Rights in the Context of Assemblies, Including Peaceful Protests' (2020) https://undocs.org/Home/Mobile?FinalSymbol=A%2FHRC%2F44%2F24&Language=E&DeviceType=Desktop&LangRequested=False consultado el 8 de octubre del 2024.

[5] Frederik Zuiderveen Borgesius, 'Discrimination, Artificial Intelligence, and Algorithmic Decision-Making' (Council of Europe, 2018) https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 consultado el 8 de octubre del 2024.

Existe una profunda asimetría entre la capacidad de aplicación y el alcance jurídico, por lo que la mayoría de los países de la Mayoría Global no pueden dar forma efectiva a los sistemas de IA instalados en sus espacios de información, incluso si los regularan. Este desequilibrio de poder perjudica la soberanía nacional y la gobernanza democrática en el ámbito digital. Incluso cuando los países elaboran normativas exhaustivas sobre IA, se enfrentan a importantes dificultades para hacerlas cumplir frente a las poderosas empresas tecnológicas multinacionales. La naturaleza transnacional de los sistemas de IA, combinada con la concentración de conocimientos técnicos y jurídicos en el Norte Global, crea una situación en la que las naciones de la Mayoría Global a menudo deben aceptar cualquier sistema y política de IA que se les imponga, independientemente de las leyes o normas sociales locales.

Las personas y la sociedad civil están al margen o excluidos en algunas parte de las fases del proceso: diseño, desarrollo, rendición de cuentas y elaboración de políticas. Necesitan acceso y un apoyo sostenido para desarrollar capacidades y conocimientos que les permitan una inclusión real. Esta exclusión perpetúa un ciclo de dependencia tecnológica y déficit democrático. Las organizaciones de la sociedad civil, que tradicionalmente desempeñan un papel crucial en la protección del interés público y la promoción de la participación democrática, a menudo carecen de los conocimientos técnicos y los recursos necesarios para participar eficazmente en la gobernanza de la IA. La complejidad de los sistemas de IA y la opacidad deliberada en su desarrollo y despliegue crean barreras sustanciales para una participación pública significativa. Esta exclusión sistemática de las voces ciudadanas significa que los sistemas de IA se desarrollan sin la aportación fundamental de las comunidades a las que más afectarán.

Esta es precisamente la razón por la que la transparencia y la explicabilidad deben ser principios cruciales en el desarrollo ético de la IA. Las personas tienen el derecho fundamental a comprender cómo los sistemas de IA afectan sus vidas y a recibir explicaciones claras sobre las decisiones automatizadas.[6] La transparencia cumple dos funciones relevantes: permite al público entender cómo funcionan los sistemas de IA y, lo que es más importante, proporciona la base necesaria para que los creadores y las plataformas se responsabilicen del impacto de sus tecnologías. Sin esa transparencia, la participación ciudadana activa y la supervisión efectiva siguen siendo imposibles, lo que refuerza aún más el desequilibrio de poder entre los desarrolladores de IA y las comunidades en las que repercuten sus sistemas.

Una de las consecuencias de la inequidad global, tal y como se ha descrito, es la presencia de la inequidad racial en el diseño,[7] su implementación,[8] el acceso a la rendición de cuentas y la

[6] Gabriela Arriagada Bruneau, Los sesgos del algoritmo: La importancia de diseñar una inteligencia artificial ética e inclusiva [The Biases of the Algorithm: The Importance of Designing an Ethical and Inclusive Artificial Intelligence] (La Pollera, 2024) https://lapollera.cl/libros/sesgos-algoritmo-ia-etica/ consultado el 28 de octubre de 2024.

[7] 'Every dataset used to train machine learning systems, whether in the context of supervised or unsupervised machine learning, whether seen to be technically biased or not, contains a worldview'. Kate Crawford, *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press 2021) 139

[8] Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' in *Proceedings of Machine Learning Research*, Conference on Fairness, Accountability, and Transparency (2018) 81:1–15 http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf consultado el 21 de octubre de 2024.

formulación de políticas. Este sesgo sistémico[9] agrava las injusticias sociales existentes y amenaza con reforzar las pautas históricas de discriminación. La falta de diversidad[10] en los equipos de desarrollo de IA, en los datos de entrenamiento y en los procesos de prueba conduce a sistemas que no sólo no abordan las desigualdades raciales existentes, sino que las refuerzan activamente. Desde los sistemas de reconocimiento facial que funcionan mal con los tonos de piel más oscuros[11] hasta los modelos lingüísticos que perpetúan estereotipos negativos, las implicaciones raciales de las actuales prácticas de desarrollo de la IA son profundas y de gran alcance. La ausencia de mecanismos eficaces de rendición de cuentas y análisis de riesgos significa que estos prejuicios a menudo pasan desapercibidos y no se abordan hasta que ya se han producido daños significativos.

En el mundo digital actual, la integridad de la información consiste fundamentalmente en dar a las personas la posibilidad de controlar su información — a qué acceden, qué consumen, cómo se presenta y cómo se evalúa. Sin embargo, nos enfrentamos a un reto crítico: un puñado de grandes empresas tecnológicas ostenta actualmente un poder sin precedentes sobre nuestro ecosistema de información, determinando lo que miles de millones de personas ven y cómo lo ven. Esta concentración de poder no es sólo una cuestión empresarial: influye directamente en el discurso democrático y a menudo lo perjudica.

La integridad de la información requiere tres elementos esenciales: transparencia en la forma en que se elabora y distribuye la información, responsabilidad por parte de quienes controlan estos sistemas y una rica pluralidad de fuentes de información fiables. Aunque los Principios Globales para la Integridad de la Información de las Naciones Unidas[12] ofrecen un marco importante — enfatizando la confianza social, los incentivos saludables, la capacitación pública, los medios de comunicación independientes y la transparencia de la investigación — su perspectiva centrada en el Norte Global requiere un análisis crítico. Desde la perspectiva del Sur Global,[13] debemos abordar retos estructurales más profundos: la soberanía digital frente a la resistencia de las plataformas a la gobernanza local, la necesidad de promover el periodismo como una práctica ética y no sólo como «medios independientes», y la reconstrucción de espacios sociales comunes en lugar de simplemente crear resiliencia frente a amenazas externas. Tenemos que ir más allá de un mundo en el que las grandes plataformas tecnológicas y sus algoritmos dominan la gestión de la información, reconociendo al mismo tiempo que un empoderamiento público significativo requiere abordar las desigualdades sociales, raciales y de género. Los Estados democráticos

[9] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown 2016) 10

[10] Forum on Information and Democracy, 'AI as a Public Good: Ensuring Democratic Control of AI in the Information Space' (febrero 2024) https://informationdemocracy.org/wp-content/uploads/2024/03/ID-AI-as-a-Public-Good-Feb-2024.pdf consultado el 10 de octubre 2024. p.24

[11] Morgan Meaker, 'This Student Is Taking On "Biased" Exam Software' (WIRED, 5 April 2023) https://www.wired.com/story/student-exam-software-bias-proctorio/ consultado el 24 de octubre 2024.

[12] United Nations, 'Global Principles for Information Integrity: Recommendations for Multi-stakeholder Action' (2023) https://www.un.org/sites/un2.un.org/files/un-global-principles-for-information-integrity-en.pdf consultado el 23 de octubre 2024.

[13] Nina Santos, 'Five Brazilian Principles for the Integrity of the Information Ecosystem' (Tech Policy Press, 2 November 2023) https://www.techpolicy.press/five-brazilian-principles-for-the-integrity-of-the-information-ecosystem/ consultado el 11 de noviembre 2024.

deben desempeñar un papel activo para evitar la concentración del mercado y garantizar un acceso equitativo a los datos y a las oportunidades de investigación, especialmente para los investigadores del Sur Global, que actualmente se enfrentan a barreras sistémicas. Esta transformación requiere no solo innovación técnica y supervisión democrática, sino también el compromiso de abordar las desigualdades estructurales que conforman nuestro ecosistema de información.

El Órgano Consultivo de Alto Nivel sobre Inteligencia Artificial ha establecido principios y recomendaciones esenciales para la gobernanza mundial de la IA, resaltando que estos principios no pueden aplicarse eficazmente sin abordar las desigualdades fundamentales, en particular entre el Norte y el Sur. Si bien la visión del Órgano Asesor de una gobernanza inclusiva y basada en los derechos proporciona una base esencial, la transformación de estos principios en realidad requiere mecanismos específicos y viables para redistribuir el poder en el ecosistema de la IA. Nuestra propuesta *IA Commons* ofrece una vía práctica para aplicar estos principios, garantizando que las voces históricamente marginadas puedan participar de forma significativa en la configuración del desarrollo y el despliegue de la IA. Este enfoque aborda directamente las preocupaciones del Consejo Consultivo sobre las lagunas de representación y los problemas de aplicación en las actuales iniciativas internacionales de gobernanza de la IA, proporcionando soluciones tangibles para construir una verdadera equidad global en la gobernanza de la IA.

Nuestra propuesta busca abordar la desigualdad global y sus múltiples componentes, en particular la desigualdad geográfica, racial y social, mediante una intervención coordinada y sistemática a múltiples niveles. Este enfoque integral reconoce que para abordar la desigualdad en la IA es necesario actuar simultáneamente en los ámbitos técnico, social y político. Exige la creación de nuevas instituciones y marcos que puedan redistribuir eficazmente el poder en el ecosistema de la IA, al tiempo que fomentan la capacidad local para una participación significativa en la gobernanza de la IA. Al abordar la naturaleza interconectada de estas desigualdades, pretendemos crear un cambio sostenible que capacite a las comunidades para dar forma a los sistemas de IA que afectan sus vidas.

No se trata simplemente de crear nuevas normativas, sino de redistribuir fundamentalmente el poder en el espacio de la información digital y garantizar que la tecnología esté al servicio de la democracia y los derechos humanos en lugar de debilitarlos. Para tal fin, proponemos los siguientes resultados prácticos para la Cumbre Mundial sobre Inteligencia Artificial de febrero de 2025, diseñados para generar un impacto inmediato y, al mismo tiempo, crear capacidad a largo plazo para la gobernanza democrática de los sistemas de inteligencia artificial.

II. IA Commons: Democratizar la IA a través del poder de la ciudadanía global

Proponemos el lanzamiento de la iniciativa *IA Commons* con cuatro pilares de aplicación. Imagina una red mundial en la que las personas, especialmente aquellos que no han tenido representatividad antes, puedan ayudar a dar forma a cómo se desarrolla y utiliza la IA. A través de cuatro programas interconectados— centros de formación en todo el Sur Global, consejos de la ciudadanía que revisen los diseños y políticas de IA, laboratorios donde la gente pueda experimentar con nuevas políticas de IA y un sistema de supervisión integral que garantice la rendición de cuentas— poniendo el poder de la IA en manos de todxs. Cada pilar desempeña un

papel vital: la Red de Laboratorios de Equidad crea capacidad, el Consejo de Diseño de la Ciudadanía permite la aportación directa de la comunidad, el Laboratorio de Innovación Política permite la experimentación segura, y el Sistema de Supervisión Multilateral garantiza que todo el proceso siga siendo responsable y transparente. No se trata únicamente de hacer que la IA sea más justa; se trata de garantizar que funcione para todas las personas y no solo para unas pocas elegidas. No se trata de una visión lejana del futuro— puede hacerse ahora para que la gente tenga herramientas que permitan que la IA fortalezca nuestras sociedades democráticas, en vez de perjudicarlas.

A. Red de Laboratorios de Equidad de IA

Piensa en la **Red de Laboratorios de Equidad de IA** como una escuela mundial para los futuros creadores de políticas de IA. Pero con una particularidad— está diseñada específicamente para potenciar las voces de las comunidades que tradicionalmente no han tenido voz en el desarrollo de la tecnología. A través de centros físicos repartidos por toda la Mayoría Global y una sólida plataforma en línea, está creando espacios en los que la gente puede adquirir experiencia práctica con sistemas de IA mientras aprende a guiar su desarrollo en la dirección correcta.

Lo que hace especial a esta red es su abordaje integral. No se trata solo de formación técnica— los participantes pasan por un programa de becas de un año en el que aprenden de todo, desde auditar los sistemas de IA para comprobar su imparcialidad hasta elaborar políticas que protejan los intereses de sus comunidades. Para 2026, la red aspira a haber formado a 1.000 nuevos líderes en políticas de IA de África, Asia, Oriente Medio y América Latina que comprendan los aspectos técnicos y sociales de la IA, creando una poderosa fuerza para el cambio positivo en el panorama mundial de la IA.

Financiación y modelo de colaboración

La sostenibilidad financiera de la Red de Laboratorios de Equidad en IA se basa en un modelo de financiación cuidadosamente elaborado por múltiples partes interesadas. En lugar de depender de una única fuente de financiación, hemos diseñado un enfoque equilibrado que distribuye los recursos y la responsabilidad entre distintos sectores. La participación de los gobiernos de los países anfitriones aporta apoyo infraestructural y legitimidad. Su inversión demuestra un compromiso con el desarrollo de la experiencia local en políticas de IA y garantiza que el programa se ajuste a los objetivos nacionales de desarrollo.

Las alianzas empresariales aportan algo más que apoyo financiero. Las principales empresas tecnológicas proporcionan recursos técnicos esenciales, oportunidades de tutoría y estudios de casos reales. Sin embargo, hemos estructurado estas asociaciones para mantener la independencia de la red y su capacidad para evaluar de forma crítica las tecnologías de IA y sus repercusiones.

Las instituciones internacionales son vitales para garantizar la relevancia global y la sostenibilidad del programa. Su participación contribuye a mantener un alto nivel de calidad y facilita el intercambio de conocimientos entre regiones. El modelo de financiación incluye mecanismos de sostenibilidad a largo plazo, como un fondo de dotación y actividades

generadoras de ingresos que apoyan el crecimiento de la red al mismo tiempo que mantienen su misión principal.

Plan de estudios y estructura de la formación

El plan de estudios de la Red de Laboratorios de Equidad en IA representa un enfoque para cerrar la brecha entre la experiencia técnica y la comprensión política en la gobernanza de la IA. En esencia, el programa reconoce que los líderes políticos eficaces en IA necesitan una comprensión global que abarque tanto las dimensiones técnicas como las sociales. Para lograrlo, hemos desarrollado un sofisticado sistema de doble vía que se adapta a la formación de las personas participantes y garantiza que todas adquieran un conjunto de competencias multidisciplinares.

El curso "Sistemas de IA y diseño de políticas," dirigido a profesionales de la política y el derecho, comienza desmitificando la tecnología de IA. Las personas participantes comienzan con experiencia práctica en programación básica y conceptos de aprendizaje automático, pasando de la comprensión teórica a la aplicación práctica. A través de laboratorios interactivos y proyectos de la vida real, aprenden a evaluar los sistemas de IA de forma crítica, a comprender sus limitaciones y a valorar su impacto social. Al final del programa, estos participantes pueden comunicarse eficazmente con equipos técnicos y tomar decisiones políticas informadas basadas en una auténtica comprensión técnica.

Las personas profesionales técnicas que entran en el programa siguen el curso "Integración de la política, los derechos humanos y la ética," transformando su experiencia técnica en conocimientos relevantes para la política. Este módulo hace hincapié en el panorama normativo, los marcos internacionales de derechos digitales y las distintas formas en que la IA afecta a diferentes comunidades. Las personas participantes aprenden a traducir sus conocimientos técnicos en recomendaciones políticas, teniendo en cuenta los diversos contextos culturales y las necesidades de la sociedad.

Ambos módulos convergen en un plan de estudios central que desarrolla aptitudes cruciales de liderazgo e incidencia. Esta experiencia compartida crea una poderosa red de profesionales capaces de reducir la brecha tradicional entre los ámbitos técnico y político.

Selección y distribución regional

El proceso de selección para la Red de Laboratorios de Equidad en IA está diseñado para construir una comunidad diversa e impactante de futuros líderes en políticas de IA. Entendiendo que las diferentes regiones se enfrentan a retos y oportunidades únicos en el desarrollo de la IA, hemos establecido un sistema de cuotas equilibrado que garantiza la representación en toda la Mayoría Global. No se trata sólo de números, sino de crear un diálogo enriquecedor entre diferentes perspectivas y experiencias.

Nuestros criterios de selección van más allá de las métricas tradicionales. Aunque la experiencia profesional es importante, valoramos especialmente a los candidatos que demuestran un profundo conocimiento de sus contextos regionales y muestran potencial para catalizar el cambio en sus comunidades. Buscamos personas capaces de conciliar la evolución mundial de la IA con las necesidades locales, teniendo en cuenta factores como su compromiso con las iniciativas

comunitarias y su capacidad para desenvolverse en complejas relaciones con las partes interesadas.

La estructura física de los centros es crucial para nuestra visión. Cada centro -ya sea en Nairobi, Yakarta o São Paulo- sirve como centro de excelencia para su región, adaptado a los contextos locales pero manteniendo los estándares globales. Estos hubs no son meros centros de formación; son incubadoras de innovación política regional en IA diseñadas para fomentar la colaboración entre las personas participantes y las partes interesadas locales.

Implementación y marco de gobernanza

La estrategia de implementación refleja nuestro compromiso con construir una institución duradera que pueda adaptarse y crecer. Nuestra estructura de gobernanza combina la supervisión global con la autonomía regional, garantizando que los programas sigan siendo pertinentes para los contextos locales, al mismo tiempo que se mantienen las normas internacionales. El Consejo Asesor Internacional desempeña un papel fundamental en la dirección estratégica, aportando diversas perspectivas de expertos técnicos, especialistas en políticas y líderes de la sociedad civil.

La garantía de calidad está integrada en todos los aspectos del programa. Las revisiones periódicas del plan de estudios, los mecanismos de evaluación de los participantes y las auditorías externas garantizan que la red siga satisfaciendo las necesidades en evolución del desarrollo de políticas de IA. La evaluación de impactos va más allá de las métricas tradicionales para evaluar cómo los participantes influyen en la política de IA en sus regiones y crean un cambio positivo en sus comunidades.

El cronograma de aplicación es ambicioso pero realista. Empezar con regiones piloto nos permite perfeccionar nuestro enfoque antes de ampliarlo. Para 2026, nuestro objetivo de formar a 1.000 líderes en políticas de IA no es solo una cuestión de números, sino de crear una masa crítica de expertos en regiones que históricamente no han sido lo suficientemente representadas en los debates sobre gobernanza mundial de la IA.

B. Consejo de Diseño de la Ciudadanía (CDC)[14]

El Consejo de Diseño de la Ciudadanía está revolucionando el diseño de la Inteligencia Artificial al dar voz a todas las personas en el proceso de desarrollo. Con importantes centros regionales situados en África, América Latina, Asia y Oriente Medio, está creando una estructura en la que las comunidades -especialmente aquellas que son ignoradas en el desarrollo tecnológico- tienen voz y voto en la construcción e implantación de sistemas de IA en sus regiones.

No se trata sólo de hacer una valoración a posteriori — el CDC involucra a las comunidades en todas las fases del desarrollo de la IA. Antes de construir cualquier sistema, evalúan su impacto cultural y las necesidades de la comunidad. Durante el desarrollo, prueban los prototipos y monitorizan sus impactos. Y después del despliegue, comprueban continuamente cómo afectan estos sistemas a la vida real de las personas. Se trata de garantizar que la IA funcione para todxs, no sólo para unos pocos expertos en tecnología.

[14] Citizen's Design Council

C. Laboratorio de Innovación en Políticas de IA (APIL)[15]

Imagina un espacio en el que diversas voces— desde responsables políticos y activistas de la sociedad civil hasta comunidades afectadas, defensores de los derechos humanos, academia y entidades privadas— puedan, en colaboración, ver y experimentar cómo afectan las políticas de IA a los derechos humanos fundamentales y a la vida cotidiana de las personas antes de su aplicación. Eso es lo que ofrece el Laboratorio de Innovación en Políticas de IA. Sus herramientas de visualización y espacios de simulación pioneros reúnen a tecnólogos, expertos en derechos humanos, líderes comunitarios, representantes empresariales y responsables políticos para transformar ideas políticas abstractas en escenarios tangibles que demuestren las repercusiones en el mundo real sobre la privacidad, la libertad de expresión, la no discriminación y otros derechos humanos esenciales.

El laboratorio reúne herramientas de alta tecnología y la elaboración de políticas centradas en los derechos humanos a través de un enfoque multilateral innovador. En sus estaciones de trabajo colaborativas y salas de política de realidad virtual, los líderes indígenas trabajan junto con los defensores de los derechos digitales, las organizaciones de base se asocian con funcionarios del gobierno, y los expertos académicos unen fuerzas con representantes de la juventud para probar cómo las diferentes políticas de IA podrían afectar a las comunidades vulnerables y las libertades fundamentales. A través de simulaciones de inmersión, este grupo diverso puede experimentar de primera mano cómo las decisiones pueden afectar al acceso a la información, la protección de la privacidad o agravar la discriminación.Se presta especial atención a los impactos interseccionales, con las comunidades afectadas liderando la conversación sobre cómo las políticas podrían afectar de manera diferente a las personas en función de su género, etnia, situación económica o ubicación geográfica. Es como tener un patio de recreo político con conciencia y sabiduría colectiva— donde las ideas pueden probarse y perfeccionarse con seguridad a través de múltiples perspectivas para garantizar que protegen y mejoran los derechos humanos antes de afectar a millones de vidas. El enfoque participativo del laboratorio garantiza que las políticas no se creen sólo para las comunidades, sino con ellas, lo que ayuda a evitar consecuencias imprevistas que puedan comprometer la dignidad humana o exacerbar las desigualdades existentes.

D. Sistema de Supervisión Multisectorial (MOS)[16]

El panorama actual de la IA presenta una paradoja crítica. Mientras empresas como OpenAI, Meta y gigantes tecnológicos regionales despliegan rápidamente sistemas de IA en todo el mundo, los mecanismos de supervisión siguen siendo dispersos y a menudo ineficaces. El estricto control de China sobre el acceso a ChatGPT frente al enfoque mayoritariamente autorregulador de EE.UU., o la normativa europea centrada en los derechos frente a los marcos emergentes de las regiones en desarrollo (limitados por los problemas de equidad descritos anteriormente). Estas disparidades resaltan la urgente necesidad de un sistema de supervisión equilibrado y adaptable que permita conciliar estos enfoques, dando prioridad a los derechos humanos y a la integridad de la información.

[15] AI Policy Innovation Lab
[16] Multi-stakeholder Oversight System

La composición de estos órganos está cuidadosamente equilibrada. Expertos técnicos trabajan junto a defensores de los derechos humanos, especialistas jurídicos colaboran con periodistas y representantes de la sociedad civil garantizan que las voces de la comunidad sigan siendo centrales en todas las decisiones. Esta diversidad no es sólo una cuestión de representación: se trata de reunir las capacidades necesarias para comprender tanto las implicaciones técnicas de los sistemas de IA como su impacto en el mundo real de las comunidades.

El sistema reconoce que la gobernanza de la IA se enfrenta a diferentes retos en las distintas regiones. En Uganda, donde el gobierno está revisando su estrategia de IA, el MOS podría proporcionar un marco para una supervisión significativa al mismo tiempo que respaldaría la innovación local. En regiones como el Sudeste Asiático, donde la localización de datos y los intereses estatales desempeñan un papel importante, el sistema podría ofrecer mecanismos flexibles que respeten la soberanía al mismo tiempo que garanticen la protección de los derechos humanos.

La capacidad de este sistema de adaptarse al par que mantiene sus principios básicos lo hace especialmente poderoso. En las regiones donde predomina la autorregulación, ofrece mecanismos de supervisión estructurados. En zonas con un fuerte control estatal, ofrece canales para la aportación de la comunidad y la protección de los derechos. Esta adaptabilidad garantiza que el sistema siga siendo pertinente y eficaz en distintos entornos normativos.

Pongamos un ejemplo práctico: Cuando una gran empresa de IA quiera desarrollar un nuevo modelo lingüístico en África Occidental, el organismo regional de supervisión evaluaría no sólo las especificaciones técnicas, sino también las implicaciones culturales, los problemas de privacidad de los datos y las posibles repercusiones en los ecosistemas de información locales. Esta evaluación no es un mero ejercicio de marcar casillas, sino una evaluación exhaustiva que puede dar lugar a modificaciones significativas o incluso a restricciones del desarrollo si es necesario.

La MOS garantiza que el desarrollo y el despliegue de la IA sigan siendo transparentes y responsables ante todas las comunidades a las que afecta, con la protección de los derechos humanos y la integridad de la información como ejes centrales. Crea un marco integral en el que los órganos regionales de supervisión, compuestos por representantes locales, defensores de los derechos humanos, periodistas, verificadores de hechos, organizaciones de medios de comunicación independientes, miembros de la sociedad civil y comunidades afectadas, tienen poder real para monitorear, evaluar e influir en la forma en que los sistemas de IA afectan tanto a los derechos humanos como a la integridad de la información en sus regiones.

La MOS supervisará los sistemas de IA implementados y en fase de pre-implementación que tengan un impacto significativo en los flujos de información y en los derechos humanos en la sociedad, centrándose específicamente en cuatro áreas críticas: (1) grandes modelos lingüísticos y sistemas generativos de IA que influyan en el discurso público y en la creación de información, (2) sistemas de recomendación y moderación de contenidos que configuren la distribución y el acceso a la información, (3) sistemas automatizados de toma de decisiones que afecten a los derechos fundamentales y a los servicios públicos, y (4) tecnologías de vigilancia y monitorización impulsadas por IA. Esta supervisión abarca la evaluación exhaustiva de las repercusiones de estos sistemas en la integridad de la información, incluido su papel en la

amplificación o mitigación de la desinformación, sus efectos en el pluralismo de los medios de comunicación y su influencia en el sesgo algorítmico en la distribución de la información. Al mismo tiempo, mantiene una rigurosa rendición de cuentas en materia de derechos humanos mediante evaluaciones de impacto obligatorias, mecanismos de reclamación vinculantes y programas de monitorización dirigidos por la comunidad. El sistema emplea un enfoque de doble vía: supervisión proactiva a través de evaluaciones previas al desarrollo y supervisión continua a través de evaluaciones posteriores a la implementación, con especial atención a los sistemas desarrollados tanto por grandes empresas tecnológicas como por entidades gubernamentales. El proceso de supervisión se basa en requisitos de documentación transparentes, audiencias públicas periódicas y mecanismos de aplicación claros, lo que garantiza que el desarrollo y el empleo de la inteligencia artificial rindan cuentas a las comunidades afectadas, protegiendo al mismo tiempo la integridad de la información y los derechos humanos.

La MOS transforma la supervisión a través de cinco enfoques clave alineados con los principios de la ONU[17] sobre la integridad de información:

1. **Fomentar la credibilidad:** Establecer mecanismos de verificación de los contenidos generados por IA y fomento de la transparencia en los sistemas algorítmicos
2. **Reestructuración de incentivos:** Hacer que las plataformas pasen de las métricas basadas en el nivel de interacción a las métricas de calidad de la información
3. **Empoderamiento público:** Apoyo a la alfabetización digital y creación de herramientas para la supervisión pública de los sistemas de IA
4. **Protección de los medios de comunicación:** Salvaguardar la independencia y la diversidad periodísticas en la era de la IA
5. **Acceso a la investigación:** Garantizar que los investigadores puedan estudiar de forma significativa el impacto de la IA en los ecosistemas de la información

A través de su enfoque integrado “derechos primero”, la MOS garantiza que los sistemas de IA respeten los derechos humanos y contribuyan a un entorno de información sano, fiable y diverso. El marco de rendición de cuentas incluye mecanismos de aplicación que van desde advertencias públicas y multas hasta restricciones operativas en caso de violaciones graves de las normas de derechos humanos o de los principios de integridad de la información.

- **Un nuevo modelo de supervisión democrática de la IA**

La MOS representa un enfoque de la supervisión de la IA situado distintamente entre los observatorios de la sociedad civil y los organismos reguladores formales. A diferencia de los tribunales, que toman decisiones vinculantes, o de los reguladores, que hacen cumplir las normas, la MOS actúa como un observatorio transparente que aclara cómo afectan los sistemas de IA a nuestro ecosistema de información y a los derechos humanos. Su poder no reside en la aplicación de las normas, sino en su capacidad para reunir pruebas, sacar a la luz patrones y permitir un discurso público informado sobre el impacto social de la IA. La función de observatorio de la MOS se ajusta a los modelos emergentes de supervisión no reglamentaria de la IA, similares al *mecanismo de certificación voluntaria para la IA de interés público* que ha

[17] United Nations, 'Global Principles for Information Integrity: Recommendations for Multi-stakeholder Action' (2023) https://www.un.org/sites/un2.un.org/files/un-global-principles-for-information-integrity-en.pdf consultado el 23 de octubre de 2024.

demostrado su eficacia en otros ámbitos. Como se destaca en la investigación del Foro sobre Información y Democracia,[18] estos mecanismos pueden ayudar a abordar las asimetrías de información entre los proveedores de IA y el público, al mismo tiempo que crean incentivos positivos para el desarrollo responsable. Al igual que los organismos de certificación que mantienen su independencia tanto de la industria como del gobierno, a la vez que fomentan la transparencia y la rendición de cuentas, la MOS puede servir como intermediario de confianza que permita una participación significativa de las partes interesadas y el debate público. Su posición como observatorio independiente le permite documentar y analizar las repercusiones sociales de los sistemas de IA, evitando a la vez los riesgos de control normativo o influencia política que pueden afectar a los organismos de supervisión más formales. Este enfoque permite a la MOS fomentar la confianza y la legitimidad a través de una evaluación rigurosa y documentación pública en lugar de a través de poderes coercitivos.

- **Legitimidad democrática a través de la estructura y el proceso**

En su esencia, la legitimidad de la MOS procede de su estructura profundamente democrática. El liderazgo rota entre representantes de diferentes regiones y grupos de interés, con estrictos límites de mandato que impiden que domine una única perspectiva. Un proceso de nombramiento transparente garantiza una representación diversa, mientras que unas políticas claras sobre conflictos de intereses y múltiples fuentes de financiación protegen contra la apropiación por parte de intereses poderosos. Las auditorías independientes periódicas de las propias operaciones de la MOS garantizan prácticas de transparencia que defiende en los sistemas de IA.

- **Fiabilidad mediante una metodología rigurosa**

Las evaluaciones de la MOS adquieren credibilidad por su rigor metodológico más que por su autoridad reguladora. Sus marcos de evaluación surgen de consultas públicas y se someten a revisión por pares, lo que garantiza que reflejen diversas perspectivas e investigaciones actuales. Antes de publicar cualquier conclusión, varios equipos independientes deben verificar los resultados y hacer pública la documentación completa de sus métodos. Este planteamiento permite obtener información fiable sobre las repercusiones sociales de los sistemas de IA, manteniendo a la vez unos límites claros entre la observación y la prescripción.

- **Medidas de protección contra "mission creep"[19] y la concentración de poder**

Para evitar que la MOS se convierta en un censor de facto o en un poder judicial paralelo, sus estatutos prohíben explícitamente las facultades de moderación de contenidos y limitan su ámbito de actuación a cuestiones sistémicas y no a casos individuales. Evaluaciones externas periódicas evalúan el cumplimiento de estas limitaciones, mientras que los informes de transparencia obligatorios detallan sus actividades y procesos de toma de decisiones. Un sólido

[18] Forum on Information and Democracy, A Voluntary Certification Mechanism for Public Interest AI: Exploring the Design and Specifications of Trustworthy Global Institutions to Govern AI (Documento de investigación, septiembre de 2024).

[19] "Mission creep" se refiere a un cambio gradual de objetivos en una misión, resultando en un compromiso alargado no planificado.

sistema de protección de los informantes fomenta la responsabilidad interna, garantizando que la MOS se mantenga fiel a su misión.

- **Supervisión colaborativa**

En lugar de trabajar de forma aislada, la MOS colabora activamente con las instituciones existentes, respetando al mismo tiempo sus distintas funciones. Aporta pruebas y conocimientos a los tribunales y a los reguladores sin intentar replicar sus funciones. Sus evaluaciones apoyan la elaboración de políticas y debate público informados sin prescribir soluciones específicas. Este enfoque colaborativo mejora la supervisión democrática de los sistemas de IA a la vez que preserva la separación de poderes, que es esencial para la gobernanza democrática.

- **Potenciar el discurso público y la comprensión**

El impacto esencial de la MOS se debe a su capacidad de aclarar cuestiones complejas para que el público las entienda. A través de reuniones públicas periódicas, sesiones de intercambio de opiniones con la comunidad y datos abiertos sobre sus operaciones, ayuda a la ciudadanía a comprender mejor las cuestiones sobre el papel de la IA en la sociedad y a comprometerse con ellas. En lugar de tomar decisiones por el público, fomenta su participación informada en debates cruciales sobre cómo la IA configura nuestro entorno informativo y los procesos democráticos.

Este enfoque cuidadosamente equilibrado garantiza que la MOS enriquezca la supervisión democrática de los sistemas de IA sin excederse en la censura o el territorio judicial. Mantener unos límites claros y, al mismo tiempo, aportar ideas cruciales refuerza las instituciones democráticas existentes en lugar de suplantarlas.

Observaciones finales

La **iniciativa *IA Commons*** representa una visión para democratizar la gobernanza de la IA a través de cuatro pilares interconectados que abordan tanto las necesidades inmediatas como los cambios estructurales a largo plazo. Combinando el desarrollo de capacidades a través de la Red de Laboratorios de Equidad de la IA, las aportaciones de la comunidad a través del Consejo de Diseño de la Ciudadanía, la experimentación política en el Laboratorio de Innovación Política de la IA y la supervisión transparente a través del Sistema de Supervisión Multisectorial, este marco crea varias vías para una participación pública significativa en la configuración del desarrollo de la IA. Es importante destacar que centra las perspectivas del Sur Global y aborda los desequilibrios de poder fundamentales en el actual ecosistema de la IA, al mismo tiempo que genera experiencia local y capacidad de toma de decisiones. Este enfoque integral reconoce que una gobernanza eficaz de la IA requiere innovación técnica y transformación social: desde abordar las desigualdades raciales y geográficas hasta garantizar la soberanía digital y reconstruir espacios sociales compartidos. El éxito de la iniciativa dependerá de un compromiso sostenido con la gobernanza inclusiva, los procesos transparentes y una auténtica redistribución del poder para garantizar que los sistemas de IA estén al servicio de los valores democráticos y los derechos humanos, en lugar de perjudicarlos.