TRANSPARENC*IA* ALGORÍTMICA

Obligaciones de Derechos Humanos en las Decisiones Automatizadas

R3D
Red en Defensa de los Derechos Digitales

TRANSPARENC*IA* ALGORÍTMICA

Obligaciones de Derechos Humanos en las Decisiones Automatizadas

R3D: Red en Defensa de los Derechos Digitales

Escrito por: Grecia Macías

Portada e interiores: Andrés Timm

Noviembre de 2025

R3D
Red en Defensa
de los Derechos Digitales

Brot
für die Welt

ÍNDICE

INTRODUCCIÓN

A Joselia de Brito, brasileña discapacitada y ex-trabajadora de la caña de azúcar, una aplicación gubernamental basada en Inteligencia Artificial (IA) denegó por error la prestación de su jubilación[1]. Según Rest of World, la aplicación la identificó erróneamente como un hombre.

Desde 2018, el Instituto Nacional de Seguridad Social (INSS) de Brasil ha implementado IA para evaluar la elegibilidad en los programas de seguridad social. Desde entonces, numerosas solicitudes, especialmente de personas en zonas remotas, han sido rechazadas debido a errores algorítmicos y procesos administrativos demasiado complejos.

En toda América Latina, los gobiernos están adoptando cada vez más sistemas predictivos automatizados de toma de decisiones, comúnmente categorizados bajo el amplio término "Inteligencia Artificial". Este tipo de IA realiza predicciones basándose en la gran cantidad de información que recibe, por lo que también se le llama "IA predictiva".

Frente a estos sistemas, que son en gran medida opacos para el público en general, las organizaciones de la sociedad civil han pedido constantemente una mayor transparencia, en concreto una transparencia significativa[2]. La transparencia está ampliamente reconocida como un principio fundamental en la gobernanza de la IA. No obstante, su invocación reiterada, sin exigencias concretas, lleva a vaciarla de contenido al usarla como una fórmula superficial sin efectos reales en la rendición de cuentas.

[1] Daros, Gabriel. "Brazil's AI-Powered Social Security App Rejected Thousands of People by Mistake." *Rest of World*, April 17, 2025. https://restofworld.org/2025/brazil-ai-social-security-app-rejected/.

[2] Derechos Digitales, *Informe comparado: IA y tecnologías predictivas en sistemas de protección social en América Latina* (2023), https://www.derechosdigitales.org/wp-content/uploads/CPC_informeComparado.pdf.

En México, el uso de sistemas predictivos en la administración pública todavía es incipiente, pero distintas reformas legislativas recientes abren la posibilidad del uso de este tipo de tecnologías, a partir de una recolección masiva de datos carece de un sistema adecuado de transparencia y rendición de cuentas.

Este informe propone reconsiderar la noción de transparencia desde un enfoque de derechos humanos, especialmente a la luz del Derecho Internacional de los Derechos Humanos (DIDH). En la primera sección, hacemos un recuento de los conceptos básicos relacionados con la IA y los sistemas predictivos, y analizamos el panorama en general de los mismos en Latinoamérica. La segunda parte desarrolla los fundamentos de los derechos humanos en relación con los sistemas automatizados, para argumentar que un marco regulatorio eficaz debe centrarse en la protección de derechos y no en enfoques basados en el riesgo. Posteriormente, identificamos las obligaciones específicas de transparencia en cada etapa del ciclo de vida de la IA, atendiendo a las necesidades de los principales actores involucrados: autoridades, auditores, personas usuarias y sociedad civil. También abordamos los límites y tensiones de la transparencia efectiva, incluyendo los de las restricciones por propiedad intelectual o por secreto industrial, así como los riesgos de sistemas que, por diseño, resultan incompatibles con los derechos humanos. Finalmente, analizamos el marco jurídico mexicano actual, sus debilidades en materia de transparencia algorítmica y presentamos estudios de caso sobre sistemas en desarrollo. Para avanzar hacia una verdadera justicia algorítmica, es urgente redefinir qué significa transparencia, a quién debe beneficiar y cómo implementarla de manera efectiva en cada fase de los sistemas predictivos.

I. SISTEMAS PREDICTIVOS EN SERVICIOS PÚBLICOS

La inteligencia artificial no tiene una mente propia; está controlada por un humano que la supervisa

En los últimos cuatro años, el término Inteligencia Artificial ha sido omnipresente. Lo hemos escuchado en redes sociales, en la boca de nuestras amistades, familiares y representantes políticos. No obstante, este término suele utilizarse para hablar de todo y nada, sin definir exactamente a qué tecnología nos referimos.

A veces, agrupamos variantes de distintos sistemas automatizados. En otras ocasiones, las discusiones refieren a sistemas generativos o a escenarios hipotéticos donde la inteligencia artificial generalizada ha ganado conciencia propia. Ese informe propone dar un paso atrás y centrarse en sistemas automatizados concretos que implican daños específicos a la población, haciendo las diferencias relevantes en cada caso.

Para hacerlo, creemos que es necesario cambiar la narrativa mistificadora que se ha creado en torno a la IA, por lo que evitaremos caer en las metáforas de una IA antropomorfizada que posee cualidades humanas[3], de las cuales los verdaderos humanos no pueden hacerse responsables.

[3] Para un análisis más completo sobre los peligros de otorgar estas atribuciones, consulte a Narayanan, Arvind y Sayash Kapoor. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton: Princeton University Press, 2024.

Tomar este camino busca evitar una mirada tecno-optimista que pasa por alto distintos riesgos que resultan de darle cualidades humanas a la Inteligencia Artificial, dentro de los que podemos señalar:

- Evitar que las personas involucradas en la creación, implementación y uso de estas herramientas sean responsables por los impactos de las tecnologías.

- Obscurecer el proceso por el cual se toman las decisiones.

- Invisibilizar las decisiones y el trabajo humano que hay detrás del algoritmo.

En sus nociones más simples, la inteligencia artificial simula el proceso de pensamiento humano a través de sistemas algorítmicos y representaciones simbólicas[4]. Por lo tanto, la IA suele clasificarse en IA débil o IA fuerte, dependiendo de qué tan similares son sus procesos a la inteligencia humana. Esto no significa que la IA más fuerte tenga una mente propia o agencia individual; simplemente se refiere al nivel de similitud con el que imitan los razonamientos humanos.

Los sistemas que adquieren, almacenan y usan conocimiento para la toma de decisiones son conocidos como sistemas basados en conocimiento[5]. Todos los sistemas de IA, desde los modelos más débiles hasta los más estrictos, manejan distintas expresiones de conocimiento o información. De manera general, estos sistemas consisten en tener una base de información con la cual entrenan al sistema para que pueda generar inferencias y proporcionar una respuesta final al usuario.

Los sistemas de aprendizaje automatizado, o machine learning (ML por sus siglas en inglés), extraen información importante de la base de conocimientos, de las personas usuarias y del

[4] Priti Srinivas Sajja, *Chapter 1: Introduction to Artificial Intelligence, in Illustrated Computational Intelligence: Examples and Applications* (Singapore: Springer, 2021), https://doi.org/10.1007/978-981-15-9589-9.

[5] Ibid.

entorno en el que se encuentran. De igual forma, estos se dividen en sistemas de aprendizaje supervisado, no supervisado e híbrido[6].

Los sistemas de ML son comúnmente usados para la IA predictiva. Esta IA predictiva usa bases de datos pequeñas o medianas para hacer predicciones basadas en modelos probabilísticos[7]. Las bases de datos generalmente se alimentan con información más específica proporcionada por la persona programadora. Para esto, se filtran y clasifican los datos según su relevancia y contenido, lo que permite realizar una predicción más específica.

Mientras tanto, **los sistemas generativos** se basan en Modelos de Gran Tamaño (LLM, por sus siglas en inglés). Estos sistemas crean contenido con texto, imágenes, video o código, y son entrenados con una gran cantidad de datos no clasificados por el programador[8]. Al utilizar principalmente información no clasificada, la IA generativa predice el valor y la relación entre cada dato mediante modelos estadísticos para generar una respuesta final. La persona programadora normalmente se encarga de revisar las respuestas de los LLMs, buscando ofrecer mayor claridad e información para perfeccionar su sistema.

Cada uno de estos sistemas se puede hacer más sofisticado para realizar tareas distintas, desde usarse para la identificación de personas en la calle hasta desarrollar redes neuronales. Así, estos sistemas pueden ser creados para un fin sumamente específico o para uno general, por lo que las implicaciones para su regulación están estrechamente relacionadas con su finalidad. En este caso, nos centraremos en el uso de sistemas predictivos con aprendizaje automatizado para la toma de decisiones en el sector público, principalmente en la administración pública.

---

[6] R3D: Red en Defensa de los Derechos Digitales. *No nos vean la cara: vigilancia en el espacio público con tecnologías de reconocimiento facial en México*, 12. Ciudad de México: R3D, mayo de 2025. *https://r3d.mx/wp-content/uploads/NNVLC_-_digital.pdf*.

[7] Jorge Garza Ulloa, "Artificial Intelligence: Predictive vs Generative vs New Mixing AI," *American Journal of Biomedical Science & Research* 22, no. 4 (May 10, 2024), *https://biomedgrid.com/fulltext/volume22/artificial-intelligence-predictive-vs-generative-vs-new-mixing-ai.002973.php*.

[8] Ibid.

Sistemas automatizados en servicios públicos

El uso de tecnologías para el despliegue de funciones relacionadas con la administración pública debe quedar enmarcado en el marco de responsabilidades jurídicas de esta última[9]. Por lo tanto, las personas servidoras públicas deben rendir cuentas sobre el uso y grado de influencia que ejercen los sistemas predictivos en la prestación de servicios públicos.

Coglianese y Lehr consideran que una IA predictiva es determinante para la toma de decisiones cuando el resultado que arroja su algoritmo es idéntico al que corresponde a la decisión administrativa final[10]. El algoritmo también determina un resultado concreto si el diseño inicial del sistema supone que se inicie una acción o se tome una decisión sin la valoración final de un ser humano. En sentido opuesto, existen sistemas predictivos que pueden ofrecer resultados que sólo representan un factor adicional, pero no determinante, para la decisión, y en los que es un ser humano quien finalmente tiene el control sobre la decisión que se toma.

Cuando los gobiernos utilizan sistemas predictivos en servicios públicos relacionados con sus funciones o en los que se utilizan recursos públicos, las personas deberían poder solicitar la rendición de cuentas de estas herramientas a través de mecanismos de transparencia[11]. Lejos de ser herramientas neutrales, estos sistemas cuentan con un diseño y despliegue que responde a una naturaleza política determinada, que se oculta en la sofisticación técnica de los mismos. Esto es importante porque, al determinar en muchas ocasiones el acceso a servicios esenciales, estos sistemas refuerzan las jerarquías sociales existentes y tienen consecuencias reales para los derechos y la vida de las personas.

[9] P. Waller y V. Weerakkody, *Digital Government: Overcoming the Systemic Failure of Transformation* (Brunel Univ. London, 2016), http://bura.brunel.ac.uk/handle/2438/12732.

[10] Cary Coglianese; David Lehr, "Transparency and Algorithmic Governance," Administrative Law Review 71, no. 1 (Winter 2019): 1-56

[11] Hogan-Doran, Dominique, SC. *Computer Says "No": Automation, Algorithms and Artificial Intelligence in Government Decision-Making.* Judicial Commission of New South Wales. https://www.judcom.nsw.gov.au/publications/benchbks/judicial_officers/computer_says_no.html, págs. 2–5.

Distintos países en Latinoamérica han empezado a implementar sistemas predictivos en los servicios públicos que otorgan. Casos como los de Argentina, Chile y Colombia ilustran distintos riesgos asociados al uso de estas tecnologías.

Argentina

En 2018, la ciudad de Salta, Argentina, implementó un algoritmo desarrollado por Microsoft para predecir embarazos adolescentes[12]. El sistema fue diseñado para predecir qué niñas y adolescentes de zonas de bajos ingresos estaban destinadas a quedar embarazadas en un plazo de cinco años. En el desarrollo de este sistema, se utilizaron datos demográficos, educativos, sanitarios y domésticos de casi 300.000 individuos para identificar a las chicas de entre 10 y 19 años que se consideraban en alto riesgo de embarazo.

Aunque se enmarcó como una intervención de salud pública, la herramienta fue entrenada y probada en un conjunto de datos sesgados, lo que resultó en una selección desproporcionada de comunidades indígenas y de bajos ingresos. Se trataba de un algoritmo diseñado de manera discriminatoria. En este caso, el sesgo no se debió principalmente a defectos técnicos o problemas de calidad de los datos, sino más bien a los prejuicios, estereotipos y suposiciones culturales de quienes participaron en su desarrollo[13]. Las organizaciones de la sociedad civil condenaron el proyecto como una forma de vigilancia reproductiva, enraizada en patrones históricos de control social y pensamiento eugenésico. El algoritmo funcionaba sin transparencia ni mecanismos de impugnación, amplificando así las narrativas patriarcales y discriminatorias bajo la apariencia de servicio público[14].

[12] Urueña, René. *Regulating the Algorithmic Welfare State in Latin America.* Max Planck Institute for Comparative Public Law & International Law (MPIL) Research Paper No. 2023-27, 20 de diciembre de 2023, p. 5. *https://ssrn.com/abstract=4670480* o *https://doi.org/10.2139/ssrn.4670480*.

[13] Smart Sebastián. *Algorithmic Discrimination in Latin American Welfare States.* Working Paper No. 24. Carr Center for Human Rights Policy, Harvard Kennedy School, 2024. *https://www.hks.harvard.edu/sites/default/files/2024-08/24_Smart_Final_01.pdf*.

[14] Urueña, *Regulating the*, 6.

Chile

El Sistema Alerta Niñez (SAN) es un sistema predictivo de IA diseñado para identificar a niños en situaciones de riesgo, basándose en más de 280 variables extraídas de bases de datos públicas[15]. Desarrollada en colaboración con organismos internacionales, esta herramienta pretendía mejorar la intervención temprana en el bienestar infantil.

Sin embargo, las bases de datos de servicios públicos que se utilizaron para entrenar a este sistema resultaron ser una muestra sesgada. Los datos excluían a los niños de las familias de altos ingresos, lo que sesgaba las evaluaciones de riesgo hacia las poblaciones pobres y marginadas. El algoritmo solía confundir pobreza con negligencia o abuso, lo que reforzaba la discriminación estructural contra las comunidades indígenas y migrantes.

A pesar de las afirmaciones oficiales sobre su precisión, el proyecto enfrentó fuertes críticas por reproducir modelos desacreditados del Norte Global e incluir la pobreza como un indicador de riesgo en la infraestructura pública. La puesta en marcha del SAN fue acompañada de un plan de evaluación de políticas en el que participaron instituciones como el Banco Mundial y el Programa de las Naciones Unidas para el Desarrollo (PNUD). Los resultados de las evaluaciones no se han hecho públicos[16].

[15] Ibid., p. 8.

[16] Derechos Digitales, *Informe comparado*, p. 37

Colombia

Por último, en respuesta a la crisis de COVID-19, el gobierno colombiano distribuyó fondos de emergencia a través del programa Ingreso Solidario, para apoyar a las poblaciones más pobres del país[17]. El programa fue elogiado por su eficacia durante el punto más álgido de la pandemia. Esta medida no requería el registro de los beneficiarios; más bien, utilizaba las bases de datos gubernamentales preexistentes y una versión actualizada del sistema nacional de clasificación de hogares, "SISBEN IV", para identificarlos.

En 2016, la metodología del SISBEN incorporó una puntuación predictiva de los ingresos futuros basada en datos de fuentes gubernamentales y del sector privado, incluidos historiales de crédito de multinacionales de calificación crediticia como Experian. Sin embargo, los criterios que respaldaban estas puntuaciones eran poco claros, lo que dificultaba una supervisión adecuada o una apelación individual.

Esto provocó que las personas con necesidades más urgentes quedaran excluidas de recibir ayuda debido a variables ocultas o registros obsoletos. Impulsado por una lógica tecnosolucionista y aplicado sin transparencia, responsabilidad ni mecanismos de reparación, el SISBEN IV hizo que las poblaciones vulnerables dependieran de un sistema de puntuación inescrutable e incuestionable.

[17] Fundación Karisma, *Experimentando con la pobreza: Algoritmos en los sistemas de ayuda social en Colombia* (2021), https://web.karisma.org.co/wp-content/uploads/download-manager-files/Experimentando%20con%20la%20pobreza.pdf.

II. LAS OBLIGACIONES DE DERECHOS HUMANOS EN EL DESARROLLO DE SISTEMAS PREDICTIVOS

La proliferación de la IA predictiva en los servicios públicos tiene profundas implicaciones para los derechos humanos, pero el diseño y el funcionamiento de estos sistemas a menudo siguen siendo opacos, dificultando la posibilidad de escrutar el grado y alcance de dichas implicaciones. Un estudio del Berkman Klein Center identificó la transparencia y la explicabilidad entre los principios más citados en los principales documentos éticos y marcos normativos sobre IA[18].

A pesar de su prominencia, estos principios se han vuelto cada vez más superficiales; se invocan con frecuencia, pero rara vez se definen o se aplican con rigor. En muchos casos, corren el riesgo de convertirse en palabras de moda, vacías y que no abordan de manera significativa los problemas estructurales de derechos humanos que plantea el despliegue de la IA.

Como sostiene Amoore en su libro *Cloud Ethics*, los principios o la regulación de la IA "deben ser capaces de plantear preguntas y hacer reivindicaciones políticas que no estén ya reconocidas en el ámbito existente de los derechos a la privacidad y las libertades de asociación y reunión[19]. Sin embargo, los sistemas predictivos con frecuencia generan resultados basados en correlaciones o inferencias novedosas que van más allá de las categorías y protecciones establecidas en los marcos de derechos humanos existentes. Por ello, es necesario que cualquier esfuerzo de

[18] Jessica Fjeld y otros, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*, Berkman Klein Ctr. Rsch. Pub. No. 2020-1 (15 de enero de 2020), https://ssrn.com/abstract=3518482.

[19] Louise Amoore, *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others* 81 (Duke Univ. Press 2020), https://doi.org/10.2307/j.ctv11g97wm.

gobernanza algorítmica esté acompañado de estándares que sean prácticos y operativos, y al mismo tiempo lo suficientemente generales para identificar estos puntos de mejora.

Marco analítico de DDHH frente a enfoques basados en el riesgo

La transparencia significativa debe integrarse en un marco más amplio de rendición de cuentas basado en los derechos humanos, que abarque todo el ciclo de vida del algoritmo. La propuesta de McGregor, Murray y Ng establece que el DIDH ofrece una estructura global para hacer operativa esta visión. Los autores sostienen[20]: "El DIDH facilita la identificación y evaluación de los daños y establece un marco claro para la rendición de cuentas basado en los derechos humanos, aplicable en todo el ciclo de vida de los algoritmos".

El DIDH establece una serie de obligaciones que los Estados deben cumplir frente al marco de derechos humanos, a la vez que diseña los mecanismos y procesos necesarios para hacer operativas estas responsabilidades. Además, este marco puede aplicarse a todo el ciclo de vida algorítmico, ofreciendo un enfoque estructurado para evaluar las responsabilidades de los diferentes actores en cada etapa. De manera complementaria, este enfoque también permite establecer obligaciones y responsabilidades para las entidades privadas, particularmente en áreas peligrosas como el bienestar y la infraestructura pública digital.

Uno de los principios fundacionales del DIDH, especialmente en el sistema interamericano, es que los Estados son los garantes de los derechos de las personas[21]. Son responsables de respetar, promover, proteger y garantizar estos derechos, mientras que los individuos son titulares de derechos que pueden exigir responsabilidades a los Estados y participar activamente en la

[20] Lorna McGregor, Daragh Murray, and Vivian Ng, "International Human Rights Law as a Framework for Algorithmic Accountability," *International and Comparative Law Quarterly* 68, no. 2 (abril de 2019): 309–343, https://doi.org/10.1017/S0020589319000046.

[21] Comisión Interamericana de Derechos Humanos, ***Políticas públicas con enfoque de derechos humanos***, OEA/Ser.L/V/II. Doc. 191, ¶ 45 (2018), https://www.oas.org/en/iachr/reports/pdfs/PublicPolicyHR.pdf.

reivindicación de esos derechos. Según el Comité de Derechos Humanos, las obligaciones de los Estados Parte en materia de derechos humanos consisten en prevenir, castigar, investigar o reparar los daños causados por su propio gobierno o por personas o entidades privadas[22].

Así, es posible afirmar que existen dos perspectivas jurídicas principales en la elaboración de políticas relacionadas con los derechos humanos y la tecnología: el enfoque basado en los derechos y el enfoque basado en los riesgos[23]. Un enfoque basado en los derechos centra la atención en las personas y les concede derechos concretos, que van desde derechos fundamentales amplios, como la intimidad y la libertad, hasta protecciones específicas, como el derecho de acceso, supresión y oposición de datos. Este modelo centra el daño en el nivel del individuo y proporciona mecanismos de reparación, autonomía y agencia.

Por el contrario, el enfoque basado en el riesgo desplaza la atención de los derechos individuales hacia la mitigación del daño sistémico[24]. Se basa en salvaguardias organizativas y técnicas, evaluaciones de riesgos y procesos de documentación, más que en recursos legales. En el contexto de la IA, esto implica priorizar la gestión de riesgos abstractos, tales como detectar y reducir sesgos, en lugar de garantizar que las personas afectadas puedan entender, impugnar u optar por no participar en las decisiones algorítmicas. Esto ilustra una limitación fundamental del modelo basado en el riesgo: carece de la fuerza normativa y de los mecanismos de rendición de cuentas que caracterizan a un marco basado en los derechos. Por lo tanto, adoptar el DIDH es fundamental para volver a enfocar la dignidad humana, la responsabilidad jurídica y las protecciones aplicables en la gobernanza de la IA predictiva.

[22] Comité de Derechos Humanos, Observación General nº 31, La índole de la obligación jurídica general impuesta a los Estados Parte en el Pacto, ¶ 8, U.N. Doc. CCPR/C/21/Rev.1/Add.13 (26 de mayo de 2004), https://www.refworld.org/legal/general/hrc/2004/en/52451.

[23] Daniel Leufer & Fanny Hidvegi, *The Pitfalls of the European Union's Risk-Based Approach to Digital Rulemaking*, 66 GERMAN L.J. 153, 159-60 (2023).

[24] Ibidem, 160.

El marco del DIDH es una guía normativa más adecuada que el enfoque basado en riesgos, para esbozar los requisitos fundamentales de cumplimiento y proteger eficazmente los derechos humanos frente a la creación y el despliegue de las tecnologías. A diferencia de las directrices éticas voluntarias sobre IA, el DIDH trata la responsabilidad algorítmica como una obligación legal, arraigada en derechos establecidos como la privacidad, la no discriminación, la libertad, el debido proceso y el acceso a la reparación. Los modelos basados en el riesgo no cumplen con las obligaciones fundamentales establecidas en el DIDH. Al tratar los derechos humanos como un factor de riesgo genérico que debe sopesarse junto con los intereses comerciales o de seguridad nacional, se socava el principio fundamental de que los derechos son inalienables y no están sujetos a transacciones.

El enfoque de riesgos normalmente sería inaceptable en un marco jurídico basado en los derechos. Sin embargo, existen ejemplos de jurisdicciones enteras donde esto es así. Por ejemplo, la Comisión Europea justifica la adopción de un modelo basado en el riesgo en la *Ley de Inteligencia Artificial de la Unión Europea* (Ley de IA) haciendo hincapié en la necesidad de evitar restricciones excesivas al comercio. Según la exposición de motivos de la Ley[25]: "Las intervenciones legales deben ser proporcionadas y ajustarse estrictamente a situaciones con preocupaciones concretas o previsibles, enmarcando el modelo como uno que permite la innovación mientras aborda los riesgos de manera selectiva".

La sociedad civil ha criticado duramente este enfoque[26]. Aunque la Ley de IA clasifica los sistemas de IA por niveles e identifica ciertos usos como no permitidos, sigue permitiendo aplicaciones profundamente invasivas, como la vigilancia biométrica masiva en contextos

[25] Id. en 165, citando la Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas sobre la inteligencia artificial (Ley sobre inteligencia artificial) y se modifican determinados actos legislativos de la Unión, exposición de motivos, en § 1.1, COM (2021) 206 final, 21 de abril de 2021.

[26] European Digital Rights (EDRi), *EU AI Act Fails to Set Gold Standard for Human Rights*, EDRi (13 de marzo de 2024), https://edri.org/our-work/eu-ai-act-fails-to-set-gold-standard-for-human-rights/.

migratorios por motivos de seguridad nacional. Esto revela una debilidad central de la ley: incluso las amenazas más graves a los derechos fundamentales no están necesariamente excluidas cuando se evalúan a la par de los intereses estatales o económicos[27].

En términos más generales, Leufer e Hidvegui sostienen que la Ley de IA carece de la flexibilidad necesaria para evaluar con precisión los daños cambiantes a los derechos humanos[28]. Un sistema de clasificación rígido es vulnerable tanto a la subinclusión como a la sobreinclusión, ya sea porque no detecta los usos realmente de alto riesgo o porque etiqueta incorrectamente las aplicaciones de bajo riesgo.

Para que sea efectiva, cualquier clasificación de riesgo debe ser adaptable y estar basada en evidencia, los ciudadanos deben tener la oportunidad de impugnar sus usos perjudiciales, y el Estado debe poder ofrecer medidas de reparación. Sin embargo, la Ley de IA, tal como está estructurada actualmente, no ofrece esa capacidad de respuesta, lo que refuerza la preocupación de que el modelo basado en el riesgo no solo es inadecuado, sino que además no se ajusta estructuralmente a la naturaleza dinámica y contextual de los daños relacionados con los derechos.

Desde las Naciones Unidas, existen propuestas regulativas tales como la de McGregor, Murray y Ng, en la que se incluyen los Principios Rectores de las Naciones Unidas sobre las Empresas y los Derechos Humanos[29] (UNGP) como bases normativas para evaluar riesgos y daños concretos.

[27] El desarrollo de armas automatizadas con inteligencia artificial es un ejemplo de cómo un concepto es inherentemente incompatible en un marco de derechos humanos. No obstante, en un marco centrado en riesgos, criterios como la seguridad nacional o la defensa se consideran junto con los derechos humanos. Para ejemplos de estos casos, consultar: Amnesty International, "Global: Amnesty International publishes an introduction to defending the rights of refugees and migrants in the digital age," Amnesty International (noticia en línea), 5 de febrero de 2024.

[28] Leufer & Hidvegi, *The Pitfalls*, 168.

[29] Consejo de Derechos Humanos de las Naciones Unidas, Principios Rectores sobre las empresas y los derechos humanos: Aplicación del marco de las Naciones Unidas para "proteger, respetar y remediar', U.N. Doc. HR/PUB/11/04 (2011).

El documento evalúa cada derecho y obligación de manera individual y, posteriormente, ofrece una serie de recomendaciones para que los Estados y las empresas prevengan, mitiguen, protejan y reparen los daños a los derechos humanos.

El DIDH es entonces clave para abordar las formas complejas y variadas en que la IA, especialmente los sistemas predictivos que la Ley de IA clasifica como de alto riesgo, puede afectar a los derechos humanos. Mientras que el *Efecto Bruselas* puede seguir influyendo en las tendencias regulatorias en América Latina, consideramos que un modelo verdaderamente eficaz para la gobernanza de la IA en la región debe estar anclado en los principios del DIDH, haciendo hincapié en los derechos exigibles y la rendición de cuentas por encima de la mera mitigación de riesgos.

Obligaciones del DIDH a lo largo del "ciclo de vida" algorítmico

Los sistemas de IA predictiva para los servicios públicos siguen el mismo ciclo de vida que la mayoría de los algoritmos predictivos. La UNESCO los ha clasificado como aquellos que abarcan: "desde la investigación, el diseño y el desarrollo hasta el despliegue y el uso, pasando por el mantenimiento, el funcionamiento, el comercio, la financiación, el seguimiento y la evaluación, la validación, el fin de uso, el desmontaje y la terminación"[30].

En cada etapa, actores estatales, privados, de la sociedad civil e investigadores académicos desempeñan roles que se entrecruzan. Estos sistemas rara vez se construyen de forma aislada. Por el contrario, surgen de complejos acuerdos público-privados en los que los límites de la responsabilidad y el acceso a la información son difusos.

Para la gobernanza de internet es vital que el sector privado conozca y cumpla con las obligaciones en materia de derechos humanos. El DIDH reconoce el *efecto horizontal* de los derechos humanos

[30] UNESCO, *Recomendación sobre la Ética de la Inteligencia Artificial* (2021). Disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000381137.

y establece un marco jurídico que describe las responsabilidades fundamentales de los actores privados frente a ellos[31]. La teoría de la eficacia horizontal de los derechos humanos reconoce que los agentes privados también poseen suficiente poder social y económico, lo que puede llevarlos a violar derechos humanos y a ser responsables por dichas violaciones[32].

Las relaciones entre particulares también se ven afectadas por el efecto expansivo de los derechos humanos. Esto supone que los derechos irradian su contenido en todas las dimensiones del ordenamiento jurídico, que incluye las relaciones entre particulares[33]. Los principios rectores sobre empresas y derechos humanos de las Naciones Unidas reflejan estas teorías. En este documento se reconoce que las empresas son "órganos especializados de la sociedad que desempeñan funciones especializadas y que deben cumplir todas las leyes aplicables y respetar los derechos humanos"[34].

El marco del DIDH es una guía normativa que esboza los requisitos básicos de cumplimiento que incluyen las responsabilidades tanto de los agentes estatales como de los privados[35]. A diferencia del marco ético de la IA, este enfatiza que la rendición de cuentas no es voluntaria, sino un requisito legal fundamentado en derechos como la privacidad, la no discriminación, la libertad, el debido proceso y el acceso a la reparación.

McGregor, Murray y Ng describen estas obligaciones del DIDH en cada fase del ciclo de vida algorítmico, desde el diseño y el desarrollo hasta el despliegue, la supervisión y el desmantelamiento, ofreciendo requisitos concretos tanto para los agentes estatales como para

[31] McGregor, Murray, y Ng, "International Human Rights Law," 312.

[32] José Juan Anzures Gurría, "La eficacia horizontal de los derechos humanos," *Cuestiones Constitucionales: Revista Mexicana de Derecho Constitucional*, no. 22 (enero–junio 2010): 13.

[33] Anzures Gurría, "La eficacia horizontal", 14.

[34] Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos, *"Principios Rectores Sobre Las Empresas Y Los Derechos Humanos"*, Nueva York y Ginebra, 2011, pág. 15.

[35] McGregor, Murray y Ng, "International Human Rights Law," 312.

los privados[36]. Igualmente, EFF creó[37] un esquema para analizar la implementación de sistemas automatizados en procesos que afectan a los derechos humanos y que incorporan los principios del sistema interamericano[38]. A continuación, haremos un repaso de estos marcos de referencia.

Fase de concepción de la medida

Antes de desarrollar cualquier modelo, los agentes públicos y privados deben evaluar si la intervención algorítmica es conceptualmente compatible con el DIDH[39]. En esta etapa, es fundamental prevenir y respetar los derechos humanos. Los modelos predictivos pueden vulnerar derechos como las garantías procesales, la igualdad o la libertad, en ciertos contextos, donde cualquier uso de algoritmos sería completamente inaceptable.

En algunas ocasiones, los Estados crean tecnología con el objetivo de resolver un problema. El uso de tecnologías con capacidad de reconocimiento facial, bases de datos biométricas y el uso de sistemas vigilancia masiva son ejemplos de tecnologías que fueron implementadas bajo una falsa premisa que prometía seguridad[40].

EFF señala que hay que limitar el problema que se quiere soluciona y responder las siguientes preguntas[41]: ¿Qué problemas busca resolver el Estado con este sistema automatizado? ¿Es un sistema automatizado la herramienta idónea, necesaria y proporcional para atender

[36] McGregor et al., "International Human Rights Law," 328.

[37] Electronic Frontier Foundation (EFF). *Human Rights Standards for the Government Use of AI in Latin America*. San Francisco: EFF, 2023. https://www.eff.org/document/human-rights-standards-government-use-ai-latin-america.

[38] EFF, *Human Rights Standards for the Government Use of AI*, 2023.

[39] EFF, *Human Rights Standards for the Government Use of AI*, 2022, 330

[40] R3D: Red en Defensa de los Derechos Digitales, *No nos vean la cara.*

[41] R3D, *No nos vean la cara*, 94-98.

este problema? ¿Qué ventajas ofrece este sistema automatizado en comparación con otros que no emplean esta tecnología? ¿Esta tecnología es confiable? ¿Qué vulnerabilidades o sesgos presenta?

De igual forma, los Estados deben analizar los grupos más vulnerables afectados por esta medida antes de desarrollar estos sistemas. Este análisis debe ser exhaustivo y estar basado en evidencia para evitar implementar una nueva política pública que cause daños irreparables desde el inicio del proceso de desarrollo. Por ejemplo, la recolección masiva y obligatoria de datos sensibles de toda la población es inaceptable debido a la dificultad para reparar el daño. Para responder a estas preguntas, los estados deben establecer mecanismos de evaluación previos y transparentes que involucren a actores relevantes de la sociedad civil, la academia y el sector técnico.

Usos inaceptables

Existen contextos en los cuales los sistemas automatizados no tienen una justificación legítima[42]. Varios de estos fueron identificados por la Ley de Inteligencia Artificial de la Unión Europea. Los siguientes son casos en los que el uso de algoritmos es fundamentalmente incompatible con los derechos:

- **Algoritmos utilizados para evadir la protección de los derechos humanos.** Por ejemplo, al inferir automáticamente o tratar de predecir categorías protegidas como la orientación sexual o la opinión política, su uso en sí constituye una violación de la dignidad y la autonomía.

- **Ausencia total de intervención humana.** Al no contar con una evaluación individualizada ni con una revisión humana, existe el riesgo de injerencias arbitrarias en derechos como la libertad, la vida familiar o el acceso a la protección social. En tales casos, el uso de la automatización es completamente inaceptable.

[42] McGregor et al., "International Human Rights Law," 335.

Fase de diseño y desarrollo

En esta etapa, los desarrolladores manejan un gran volumen de información sensible y toman decisiones cruciales respecto a la arquitectura de los sistemas automatizados. Igualmente, este es el momento en el que se crean los protocolos relacionados con la interacción entre funcionarios públicos y este sistema. En consecuencia, los Estados y los desarrolladores deben realizar evaluaciones de impacto sobre los derechos humanos (EIDH) que analicen los posibles daños a los derechos antes de poner la IA a disposición del público.

Recientemente, algunas normativas exigían evaluaciones de impacto en la protección de datos personales[43]. No obstante, esto deja de lado otras afectaciones a derechos como la igualdad, la no discriminación y el derecho al debido proceso. Estas evaluaciones se correlacionan con las obligaciones de proteger y garantizar los derechos humanos de la ciudadanía. Por lo tanto, los Estados deben asegurarse de que los sistemas automatizados cuenten con mecanismos de mitigación de riesgos que puedan afectar los derechos humanos, especialmente los derechos de los grupos en situación de mayor vulnerabilidad.

Las pruebas de estos sistemas automatizados deben realizarse de manera constante para garantizar un nivel básico de protección a la ciudadanía. Por lo tanto, las etapas de prueba no pueden incluir el despliegue de esta tecnología en casos reales hasta que se implemente un mecanismo mínimo de protección.

Fase de despliegue

Esta etapa comienza cuando el sistema automatizado interactúa por primera vez con las personas usuarias en contextos reales. Aunque muchos gobiernos optan por una implementación gradual y en ubicaciones limitadas, incluso estos despliegues iniciales deben contemplar garantías adecuadas. Es fundamental que el despliegue incluya mecanismos explicables, acceso a recursos de apelación y supervisión independiente.

[43] En secciones posteriores, analizamos el marco jurídico de protección de datos personales en México.

Es importante no confundir esta etapa con las pruebas técnicas propias del desarrollo. Durante la fase de desarrollo pueden utilizarse datos personales, pero el sistema aún no presta servicios directamente a la ciudadanía. En cambio, la etapa de despliegue sí involucra interacciones reales con personas fuera de ambientes controlados, lo que activa nuevos riesgos para los derechos humanos. Mientras que en el desarrollo el riesgo principal puede ser la exposición de información sensible, en el despliegue puede traducirse en la negación injustificada de beneficios públicos debido a fallos no anticipados del sistema. El hecho de que un sistema no esté implementado al 100% en todas las localidades no exime a las autoridades de su responsabilidad: no pueden justificarse diciendo que "aún están probando" el sistema.

Finalmente, en esta etapa es de vital importancia comenzar a capacitar al personal que utilizará el sistema automatizado. Los sistemas que incluyen a un "humano en el ciclo" (*human-in-the-loop*) no constituyen una salvaguarda automática. Su efectividad depende de que las personas responsables de su uso puedan comprender, cuestionar y, cuando sea necesario, contradecir las decisiones algorítmicas.

Fase de supervisión y uso

Las obligaciones del DIDH exigen que los actores supervisen continuamente los sistemas desplegados y los revisen a medida que surgen nuevos riesgos[44]. Cuando los sistemas aprenden o evolucionan en tiempo real, la obligación de reevaluar el riesgo se vuelve constante. Los equipos de auditoría interna deben complementarse con organismos de supervisión externos e independientes que tengan la autoridad para suspender o restringir su uso si es necesario.

Fase de suspensión

Por último, cuando se comprueba que los sistemas producen un daño injustificable, el marco del DIDH exige su interrupción[45]. Además, las personas afectadas deben poder acceder a soluciones efectivas, incluyendo el acceso a la reparación y garantías de no repetición.

[44] McGregor et al., "International Human Rights Law," 331.

[45] McGregor et al., "International Human Rights Law," 332.

La transparencia en las etapas de desarrollo de los sistemas predictivos

La transparencia de la IA es un concepto polifacético y con diferentes niveles. La noción de transparencia algorítmica es controvertida, especialmente en el contexto de la toma de decisiones gubernamentales donde los servidores públicos tienen la responsabilidad de justificar sus acciones. No obstante, el DIDH y el Sistema Interamericano de Derechos Humanos (SIDH) son guías apropiadas para desglosar las distintas obligaciones fundamentales para proteger el derecho a la información.

Obligaciones de transparencia del SIDH

Dentro del SIDH, existen obligaciones especiales relacionadas con el principio de máxima divulgación. Este principio ordena crear un régimen jurídico en el cual la transparencia y el derecho de acceso a la información sean la regla general, sometida a estrictas y limitadas excepciones. De igual forma, los Estados parte deben someterse a un escrutinio más riguroso al establecer limitaciones al acceso a la información. Por lo tanto, en caso de que se niegue la información, el Estado debe fundamentar y motivar el rechazo de la solicitud. Al mismo tiempo, el derecho de acceso a la información está relacionado con la libertad de expresión[46], por lo que las limitaciones que se impongan al acceso a la información también deben cumplir con los estándares de excepcionalidad, legalidad, necesidad y proporcionalidad.

El Estado tiene la carga de la prueba para demostrar por qué la información reservada o clasificada como confidencial cumple con este escrutinio. Al hacerlo, los gobiernos no deben usar la justificación de seguridad nacional como una categoría absoluta y sin límites, que sirva para restringir la información de manera categórica y sin explicación, sino enmarcarla en una interpretación democrática en sí misma.

Por ello, uno de los estándares más consensuados dentro del Sistema Interamericano en este aspecto señala que: “La necesidad de realizar un juicio de proporcionalidad estricta implica

[46] Artículo 13 de la Convención Americana de Derechos Humanos.

que ninguna información puede ser excluida de antemano del control público, simplemente por estar en poder de un organismo de seguridad nacional, por estar relacionada con esta materia o por encajar en una determinada categoría de información[47]".

Fundamentos de la Transparencia algorítmica basada en DIDH

Existen sistemas predictivos que funcionan como una "caja negra". Dentro de la informática, una caja negra suele referirse a sistemas en los que sólo son observables las entradas y salidas, mientras que el funcionamiento interno permanece oculto.

Como explica Dominique Hogan-Doran SC[48]: "En los sistemas de toma de decisiones completamente automatizados, tanto las entradas como las salidas también están encapsuladas en la caja negra. Las fuentes de información que representan los insumos no son observables. Sin embargo, para el estado de derecho es fundamental que los ciudadanos tengan apertura y la posibilidad de formar su propia opinión y discrepar si no están de acuerdo con el ejercicio de los poderes gubernamentales. El acceso a la información sobre las normas detalladas del sistema debe hacerse público".

Como mínimo, la transparencia de la IA debe incluir información sobre los insumos, los procesos y los resultados. Esto no quita que deban evaluarse los distintos escenarios y sistemas sobre los que ésta se exige; existen escenarios donde la estricta transparencia no es ideal al considerar aspectos relacionados con la privacidad de los usuarios o elementos que efectivamente afecten a la seguridad pública[49].

[47] Comisión Interamericana de Derechos Humanos (CIDH), *Derecho a la información y seguridad nacional*, Relatoría Especial para la Libertad de Expresión, OEA/Ser.L/V/II., CIDH/RELE/INF.3/17 (2013), https://www.oas.org/es/cidh/expresion/informes/DerechoInformacionSeguridadNacional.pdf.

[48] Hogan-Doran, *Computer Says "No"*, 32.

[49] Joshua A. Kroll, "Accountability in Computer Systems," en *The Oxford Handbook of the Ethics of AI*, eds. Markus D. Dubber, Frank Pasquale, y Sunit Das (Oxford: Oxford University Press, 2020), https://ssrn.com/abstract=3608468; Karthik Haresamudram, Staffan Larsson y Fredrik Heintz, *Three Levels of AI Transparency*, 56 Computer 93 (2023).

Sin embargo, en discusiones sobre la aplicación de este principio, suele creerse que la transparencia significativa se alcanza con el mero hecho de tener acceso al código fuente de un sistema, cuando ésta va más allá de este mero hecho. La transparencia significativa requiere información práctica y contextualizada para las distintas partes interesadas: reguladores, personas afectadas, auditores técnicos y sociedad civil.

Algunos académicos consideran que existen tres niveles principales de transparencia de la IA[50]: algorítmica, de interacción y social:

- **Transparencia algorítmica.** Se refiere a hacer comprensibles la lógica interna, los datos de entrenamiento y las reglas de decisión de un sistema, principalmente (aunque no únicamente) para expertos técnicos y auditores.

- **Transparencia de interacción.** Se centra en cómo se relacionan los usuarios con los sistemas de IA, haciendo hincapié en las interfaces que permiten a las personas comprender y, en caso necesario, rebatir las decisiones algorítmicas.

- **Transparencia social.** Se refiere al contexto institucional, jurídico y cultural más amplio en el que operan estos sistemas. Es esencial para las organizaciones de la sociedad civil, los periodistas y los organismos de supervisión que pretenden evaluar las repercusiones sistémicas y exigir responsabilidades a los actores.

Las expectativas de transparencia varían según los grupos de interesados. Las comunidades técnicas pueden dar prioridad al acceso a la información sobre la arquitectura del modelo, los datos de formación y los parámetros de rendimiento. Los académicos, las instituciones públicas o las organizaciones de la sociedad civil, por el contrario, suelen centrarse en la procedencia de los datos, el cumplimiento legal y los mecanismos de rendición de cuentas y supervisión[51].

[50] Karthik Haresamudram, Staffan Larsson y Fredrik Heintz, *Three Levels of AI Transparency*, 56 Computer 93 (2023), https://doi.org/10.1109/MC.2022.3213181.

[51] Derechos Digitales, *Informe Comparado*, 26.

Aunque los distintos tipos de transparencia refieren de manera general a sujetos distintos, lo ideal es que las distintas categorías estén abiertas a los sujetos interesadas en ellas.

Las personas usuarias son una parte fundamental, pero a menudo olvidada, del ecosistema de la transparencia. Los desarrolladores suelen pasar por alto las necesidades de las personas directamente afectadas por las decisiones automatizadas. Para que la transparencia tenga sentido, estas personas deben disponer de herramientas e información para comprender qué sistema se está utilizando, cómo funciona y cómo afecta sus derechos y su vida cotidiana. Sin esta comprensión, cualquier pretensión de consentimiento informado sigue siendo ilusoria.

Es aquí donde el *principio de explicabilidad* entra en juego. En la gobernanza de sistemas de IA, no es suficiente que se describan las conclusiones a las que llegó un sistema predictivo, sino que también es necesario saber por qué se llegó a esa conclusión[52]. Estas explicaciones son especialmente importantes con sistemas predictivos que interactúan con espacios y variables dinámicas. Las justificaciones también deben realizarse en función de la persona y del momento en que se utiliza un sistema predictivo. Por ejemplo, un sistema predictivo utilizado para otorgar un préstamo debe poder explicar a las personas usuarias las razones específicas por las que se rechazó ese préstamo y cómo se llegó a esa decisión. Asimismo, debe proporcionar explicaciones a los desarrolladores que realicen las auditorías para que comprendan técnicamente por qué se tomó esa decisión.

El principio de explicabilidad está directamente relacionado con la obligación de los actores gubernamentales de fundamentar y motivar sus decisiones que afectan a la ciudadanía. Ya sea considerado como una modalidad del principio de máxima publicidad o como un fundamento básico del derecho al debido proceso.

En la siguiente sección se describen las obligaciones específicas de transparencia y explicabilidad que deben aplicarse en cada fase del ciclo de vida de la IA predictiva, tal como se derivan del DIDH.

[52] David Danks, "Governance via Explainability," en *The Oxford Handbook of AI Governance*, Justin B. Bullock, capítulo 9 (Oxford: Oxford University Press, 2022), 185.

La explicabilidad según la fase del ciclo de vida

Existen diferentes grados de oscuridad para la caja negra algorítmica. No todos los sistemas son igualmente opacos. En algunos casos, la decisión automatizada puede trazarse de forma tan sencilla como siguiendo un diagrama de flujo. Sin embargo, incluso si las entradas y salidas son observables, esto a menudo no proporciona una interpretabilidad real[53]. Los modelos de aprendizaje profundo, por ejemplo, son inherentemente inexplicables, por lo que podrían ser inherentemente incompatibles con las obligaciones del DIDH para la IA predictiva, las cuales requieren diferentes niveles de transparencia en cada etapa del ciclo.

Los algoritmos de predicción policial que carecen de una revelación significativa deberían prohibirse para uso policial, al igual que los algoritmos inexplicables no deberían determinar la elegibilidad para la asistencia social. Estos sistemas están vinculados a las funciones gubernamentales y afectan a servicios públicos cruciales para los ciudadanos. La transparencia es esencial desde el principio para garantizar la rendición de cuentas y evitar el despilfarro de inversiones en IA predictiva que no sea transparente.

En escenarios de alto riesgo, como sucede con los sistemas de bienestar público, los Estados deberían favorecer los modelos intrínsecamente interpretables. Cynthia Rudin argumenta que "los modelos interpretables, que proporcionan un equivalente técnico, pero posiblemente una alternativa más ética a los modelos de caja negra, son diferentes: se limitan a proporcionar una mejor comprensión de cómo se realizan las predicciones[54]".

Primero, las autoridades deben analizar los objetivos que buscan lograr con los sistemas predictivos. Este análisis debe realizarse tras una evaluación previa del impacto en los derechos humanos. Para ello, las autoridades deben convocar a las diferentes partes interesadas para que puedan expresar su opinión experta sobre estos riesgos. Asimismo, el análisis de impacto en

[53] Jens Christian Bjerring, Jakob Mainz y Lauritz Munch, Deep Learning Models and the Limits of Explainable Artificial Intelligence, 4 Asian J. Phil. 22 (2025), https://doi.org/10.1007/s44204-024-00238-8.

[54] Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 Nat. Mach. Intell. 206 (2019), https://doi.org/10.1038/s42256-019-0048-x.

derechos humanos debe ser integral y no limitarse únicamente a una evaluación del impacto en la privacidad. Por ejemplo, debe analizarse el impacto ambiental que generaría el uso de sistemas predictivos que requieran centros de datos masivos para funcionar. El derecho a la igualdad y no discriminación exige también que las autoridades evalúen que el sistema no genere tratos o efectos discriminatorios.

Fase de diseño y desarrollo

Esta fase suele ser la menos transparente, pero es una de las más importantes. El diseño y el desarrollo suelen realizarse a puerta cerrada, bajo la dirección de programadores y científicos de datos, sin una participación significativa de las comunidades afectadas. Esta cámara de eco técnica puede amplificar los prejuicios y pasar por alto las realidades vividas por los grupos vulnerables[55]. Tanto si el Estado desarrolla la IA predictiva internamente como si la subcontrata a agentes privados, la transparencia en esta fase debe ir más allá de simplemente compartir el código fuente con algunas partes interesadas.

Un enfoque técnico sólido de la transparencia debe incluir documentación contextual: reglas de decisión, parámetros de rendimiento, protocolos de validación y prueba, así como auditorías de sesgo. Además, también es la fase en la que los algoritmos se entrenan con grandes cantidades de datos, muchos de ellos personales y sensibles. Por lo cual, son buenas prácticas para promover la transparencia tanto crear fichas que expliquen la procedencia[56] y el tratamiento de la información contenida en bases de datos, como generar reportes que expliquen el funcionamiento de modelo[57].

[55] Kate Crawford y Ryan Calo, *There Is a Blind Spot in AI Research*, 538 Nature 311 (2016), https://doi.org/10.1038/538311a.

[56] Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III y Kate Crawford. "Datasheets for Datasets." *arXiv* (preprint), publicado el 23 de marzo de 2018, última versión revisada el 1 de diciembre de 2021. https://doi.org/10.48550/arXiv.1803.09010.

[57] Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, y Timnit Gebru. "Model Cards for Model Reporting." *Proceedings of the Conference on Fairness, Accountability, and Transparency* (*FAT* '19) *, enero de 2019. https://doi.org/10.1145/3287560.3287596. researchgate.net+8

La documentación sobre la procedencia de los datos también está relacionada con la transparencia social. Los desarrolladores deben revelar qué tipos de datos se utilizan durante el entrenamiento, si se obtuvo el consentimiento y cómo se seleccionaron y procesaron las fuentes de datos[58]. Este tipo de documentación ayuda a mejorar los procesos de implementación y evaluación, previniendo afectaciones a los derechos humanos y fortaleciendo la gobernanza de los sistemas de IA[59].

Incluso cuando el modelo no se centra explícitamente en características protegidas, las obligaciones de transparencia exigen la divulgación de las estrategias de mitigación de sesgos, los resultados de la validación y las limitaciones conocidas del modelo. En algunas ocasiones, la transparencia del código en sí misma no revelará muchos de los sesgos discriminatorios presentes en las bases de datos. Un sector de la academia[60] considera que, en la fase de desarrollo, también es necesario aplicar factores correctivos al sistema que se basa explícitamente en una categoría sospechosa. En otras palabras, "vacunar" a los sistemas con variables que utilizan directamente categorías discriminatorias desde una etapa inicial, para que eviten replicarlas en la fase de operación.

Otra consideración importante en esta fase es cómo se gestiona la recopilación de datos. La transparencia no exige que los desarrolladores o los gobiernos revelen públicamente toda la información personal, especialmente los datos sensibles[61]. Una salvaguarda adecuada consiste en anonimizar la información y en intercambiar datos únicamente con terceros autorizados

[58] A.G. Res. 78/265, en 6, U.N. GAOR, 78º Sess., Agenda Item 13, U.N. Doc. A/RES/78/265 (1 de abril de 2024).

[59] Bogen, Miranda, y Amy Winecoff. "Best Practices in AI Documentation: The Imperative of Evidence from Practice." *Center for Democracy & Technology*, 25 de julio de 2024.

[60] Christopher S. Yoo, "Beyond Algorithmic Disclosure for AI," Columbia Science and Technology Law Review 25 (2024): 317, *https://scholarship.law.upenn.edu/faculty_articles/426/*. Deborah Hellman, "Measuring Algorithmic Fairness," Virginia Law Review 106, no. 4 (2020): 811–866, *https://virginialawreview.org/wp-content/uploads/2020/06/Hellman_Book.pdf*.

[61] Alejandro Barredo Arrieta et al., *Inteligencia artificial explicable (XAI): Conceptos, taxonomías, oportunidades y retos hacia una IA responsable*, 58 Inf. Fusion 82, 43 (2020), *https://doi.org/10.1016/j.inffus.2019.12.012*.

previamente por las personas titulares. Sin embargo, esto plantea una cuestión más profunda y preocupante: ¿debería la IA predictiva manejar datos tan sensibles en primer lugar? Se trata de una preocupación que la transparencia por sí sola no puede resolver. Por ejemplo, en el caso del SISBEN, el sistema incorporó información financiera sin el consentimiento de la persona, lo que revela un fallo fundamental en su diseño y en su propósito.

Otro caso que ejemplifica este dilema es el uso de algoritmos de puntuación basados en riesgo para determinar si una persona es elegible para un apoyo monetario[62]. El Relator Especial de Naciones Unidas en materia de Pobreza Extrema advirtió sobre los peligros de estos sistemas[63]. Entre ellos se encuentran los errores que pueden tener los sistemas predictivos de riesgo, la opacidad con la que funcionan estas tecnologías y cómo, en general, estos sistemas empeoran la desigualdad y la discriminación preexistentes.

Aunque algunos estados intentan mitigar la opacidad mediante sistemas de código abierto u ofrecer resúmenes generales de estas herramientas, esta divulgación suele ser insuficiente para que las poblaciones afectadas comprendan o impugnen las decisiones. Principalmente, porque no abordan el problema de raíz: la clasificación ilegal de personas basada en categorías protegidas.

Fase de despliegue

Dado que esta fase implica que la IA predictiva interactúe con personas usuarias fuera de un entorno controlado, las obligaciones de transparencia deben centrarse en la instalación del sistema, los protocolos en desarrollo, los errores que puedan surgir durante el despliegue y los mecanismos para apelar las decisiones.

[62] Burgess, Matt, Evaline Schot, y Gabriel Geiger. “This Algorithm Could Ruin Your Life.” Wired, 6 de marzo de 2023, *https://www.wired.com/story/welfare-algorithms-discrimination/*

[63] Relator Especial de la ONU sobre la extrema pobreza y los derechos humanos. *Informe presentado al Consejo de Derechos Humanos, A/74/493*, párrafo 28. Nueva York: Naciones Unidas, 2019.

Esta fase es igualmente importante para determinar la cadena de responsabilidad de los servidores públicos frente al abuso de esta herramienta. Como parte del principio de máxima publicidad, los gobiernos deberán crear sistemas que especifiquen quiénes son las personas autorizadas para utilizarlos y delimiten los procesos en los que serán empleados. Estos protocolos pueden revisarse en conjunto con las partes interesadas como una forma de promover la transparencia social.

Para el sector técnico, una transparencia significativa requiere interfaces explicativas para el usuario, documentación detallada de las interacciones entre el ser humano y la IA, y datos empíricos sobre el rendimiento del sistema en condiciones reales.

Los gobiernos también deben garantizar la participación activa de diversas partes interesadas durante el despliegue para identificar cualquier daño imprevisto a los derechos humanos que pueda no haber surgido en pruebas anteriores. Las organizaciones de la sociedad civil, los expertos en políticas públicas y los académicos se preocuparán especialmente por los grupos a los que está dirigido el sistema y si existe un plan de implementación gradual.

Finalmente, es importante que las partes interesadas tengan la oportunidad de impugnar cualquier daño que surja durante esta fase de despliegue. El gobierno debe mantener un registro de cualquier impugnación o queja relacionada con estos sistemas y hacerlo accesible para que auditores externos puedan ofrecer comentarios que prevengan futuras afectaciones a los derechos humanos.

Fase de seguimiento y uso

Una vez desplegada la IA predictiva, los usuarios necesitarán más información sobre el sistema antes de usarlo o inscribirse en él. Una referencia podrían ser marcos como el Reglamento General de Protección de Datos (RGPD) de la UE, en el que las personas tienen derecho a obtener "información significativa" sobre la lógica que subyace a las decisiones automatizadas[64].

[64] Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (Reglamento general de protección de datos), arts. 13(2)(f), 14(2)(g), 2016 O.J. (L 119) 1.

Sin embargo, la transparencia significativa va más allá de la divulgación. Requiere que la información sea procesable para diferentes audiencias: individuos afectados, sociedad civil, auditores técnicos y reguladores.

Recientemente, la Corte Constitucional de Colombia resolvió[65] un caso emblemático sobre transparencia algorítmica para las personas usuarias. La Corte colombiana examinó una tutela interpuesta por un ciudadano contra la Agencia Digital Nacional, el Instituto Nacional de Salud y el Ministerio de Salud y Protección Social. El peticionario alegó una violación de su derecho fundamental de acceso a la información pública después de que las autoridades se negaran a revelar el código fuente de *CoronApp*, una aplicación móvil desarrollada por el Estado y creada en respuesta a la pandemia de COVID-19.

Las autoridades demandadas alegaron que la información solicitada estaba exenta de divulgación en virtud de las disposiciones sobre confidencialidad. Argumentaron que la divulgación del código fuente pondría en peligro la privacidad de los datos personales recogidos por la aplicación, exponiéndolos potencialmente a un uso indebido por parte de actores maliciosos. Además, alegaron que la publicación del código pondría en peligro la eficacia y la seguridad de las medidas de salud pública aplicadas durante la pandemia de COVID-19. Por último, las autoridades invocaron la protección de los derechos de autor, afirmando que la divulgación infringiría los derechos de propiedad intelectual de los desarrolladores de la aplicación.

El tribunal dictaminó que la información solicitada por el demandante debía divulgarse y, al hacerlo, aclaró el significado de la transparencia algorítmica. Este concepto busca asegurar que el público comprenda cómo los sistemas automatizados de toma de decisiones procesan los datos que recopilan y cómo toman decisiones que afectan la vida de las personas. Es un principio con una finalidad constitucional: democratizar el funcionamiento interno de esos sistemas para que sean comprensibles para quienes se ven afectados por su aplicación y funcionamiento.

Como explicamos anteriormente, hacer público el código fuente, aunque puede ser un gran paso para la transparencia algorítmica, no es suficiente para alcanzar el estándar de explicabilidad

[65] Corte Constitucional de Colombia, Sentencia T-067/25 (2025). Disponible en: *https://www.corteconstitucional.gov.co/relatoria/2025/t-067-25.htm*.

ante terceros. Por ejemplo, una persona sin conocimientos técnicos para interpretar el código fuente tendrá las mismas dudas sobre el tratamiento de sus datos por la app.

Por último, otra medida importante de transparencia social en esta fase es habilitar un mecanismo activo de retroalimentación y un sistema de auditores para promover la supervisión activa del uso de la IA predictiva. Estas obligaciones no solo incluirán evaluaciones técnicas sobre los índices de error, sino también la detección de posibles abusos humanos o mal uso de la IA predictiva.

Fase de suspensión o finalización

Los sistemas deben retirarse progresivamente si las evaluaciones revelan que no cumplen con las normas de derechos humanos. En este contexto, la transparencia requiere una justificación clara de la retirada, la divulgación pública de los perjuicios detectados y la implementación de un proceso de apelación abierto para garantizar que los afectados puedan acceder a una reparación.

Obstáculos y límites para la transparencia efectiva

En los últimos años, a raíz de escándalos públicos, han existido iniciativas empresariales que pretenden promover políticas de ética y transparencia. Sin embargo, éstas se han centrado en crear procesos que distraen o legitiman conductas que ponen en riesgo los derechos humanos. Generalmente, las empresas (y los gobiernos) invocan el lenguaje de la ética de la IA y la apertura mientras ocultan información crítica de sus sistemas, tal como los datos de formación o los métodos de evaluación interna[66], centrándose en principios de alto nivel más que en obligaciones vinculantes[67] (permitiendo a las empresas asumir posturas éticas sin ninguna

[66] Diya, Sabhanaz Rashid. *Applying International Human Rights Principles for AI Governance.* CIGI Policy Brief No. 196, 12 de febrero de 2025. *https://www.cigionline.org/publications/applying-international-human-rights-principles-for-ai-governance/*.

[67] Ibidem.

responsabilidad)[68]. Todo esto muestra una postura en la que se busca legitimar sistemas inaceptables en sí mismos por el mero hecho de hacerlos visibles[69].

Teniendo esto en mente, es importante recordar que la transparencia no es un fin en sí mismo, sino una herramienta para redistribuir el poder y fomentar la rendición de cuentas, tanto de las empresas como de los gobiernos que utilizan sistemas automatizados. En los apartados anteriores, esbozamos ciertos límites naturales de la transparencia algorítmica que deben ser considerados al realizar acciones sobre sistemas automatizados. Ahora analizaremos dos circunstancias importantes que generan tensiones con la transparencia algorítmica.

Límites y peligros de las evaluaciones de sistemas algorítmicos

Una de las principales recomendaciones en materia de transparencia algorítmica es realizar evaluaciones de impacto o auditorías de los sistemas. Los análisis de impacto previo normalmente buscan anticipar los impactos del sistema o para decidir cómo limitar su uso futuro[70]. Es un análisis preventivo que muchas veces realizan los desarrolladores de manera individual. Las auditorías normalmente se realizan cuando los sistemas ya están en la etapa de desarrollo y se tiene un objetivo específico de acceder a información y emitir un juicio a partir del análisis de esta información[71].

[68] Ibidem.

[69] Zalnieriute, Monika. "Transparency Washing in the Digital Age: A Corporate Agenda of Procedural Fetishism." *Canadian Journal of Law & Technology* 32 (2024): 1. https://cal.library.utoronto.ca/index.php/cal/article/view/36284.

[70] Selbst, Andrew D. "An Institutional View of Algorithmic Impact Assessments." *Harvard Journal of Law & Technology* 35, no. 1 (Fall 2021): 151–204. https://jolt.law.harvard.edu/assets/articlePDFs/v35/35HarvJLTech151.pdf.

[71] Raji, Inioluwa Deborah. "The Anatomy of AI Audits: Form, Process, and Consequences." In *The Oxford Handbook of AI Governance*, edited by Justin B. Bullock et al. Oxford Handbooks. Oxford: Oxford University Press, 2024. Online edition, Oxford Academic, February 14, 2022. https://doi.org/10.1093/oxfordhb/9780197579329.013.28.

Normalmente, una auditoría de IA busca corroborar si el actual funcionamiento del sistema corresponde a los estándares y objetivos iniciales del programa. En general, las auditorías no solo son una forma de acceder a la información, sino que también constituyen un mecanismo de rendición de cuentas. La necesidad de realizar auditorías con partes interesadas independientes no es nueva. Los análisis independientes también son vistos en áreas financieras, ambientales o en la evaluación de políticas públicas[72].

Aunque es cierto que estos dos tipos de evaluaciones han funcionado como un mecanismo de rendición de cuentas en la gobernanza algorítmica, ayudando a prevenir o mitigar riesgos, el diseño e implementación de estas evaluaciones no es una tarea sencilla ni resuelve por sí misma los problemas que hemos advertido. En general, estos mecanismos generan confianza, pero también producen una falsa sensación de seguridad; si las evaluaciones se realizan de manera incorrecta, podrían llevar a aprobar IAs peligrosas para la población[73].

Por ello, consideramos que hay cuatro factores que deben tenerse presentes al momento de realizar este tipo de análisis, tanto para el momento de la planeación de lo mismos como para cuando se evalúen sus resultados:

- **No existe un modelo único para estas evaluaciones.** Aunque es necesario contar con evidencia para analizar los sistemas automatizados, necesitamos información relevante para cada caso particular. Por ejemplo, una auditoría a un sistema predictivo tendrá indicadores diferentes a los de una auditoría a un LLM. Igualmente, el modelo depende del objetivo de investigación. Una auditoría en materia de privacidad requiere un nivel de transparencia diferente al de una evaluación técnica para identificar sesgos.

- **La independencia de los evaluadores es fundamental.** Es común que realizar análisis independientes de impacto previos sea más difícil, porque los gobiernos y las empresas desarrolladoras seleccionan personal interno y utilizan indicadores de

[72] Ibidem.

[73] Goodman, Ellen P., y Julia Trehu. "AI Audit Washing and Accountability." *SSRN Scholarly Paper*, 22 de septiembre de 2022. *https://ssrn.com/abstract=4227350 o https://doi.org/10.2139/ssrn.4227350*.

evaluación que reflejan la mirada interna del desarrollo se sus sistemas. Al hacerlo, suelen amparar estas acciones de ostracismo en argumentos relacionados en el interés por el secreto industrial de sus productos. Y aunque es cierto que contar con auditores externos puede resolver problemas de independencia y opacidad, estos muchas veces enfrentan diversos obstáculos para acceder a información relevante. Por ejemplo, en el Reglamento de Servicios Digitales[74], se requiere que las plataformas de gran tamaño (VLOP, por sus siglas en inglés) realicen auditorías de sus sistemas algorítmicos con expertos independientes que hayan sido previamente aprobados por dichas plataformas, pero los requisitos para obtener el estatus de investigadores aprobados son estrictos y están sujetos a la burocracia interna de cada plataforma.

» **Las evaluaciones cosméticas o "audit washing" son la regla.** Una auditoría bien realizada puede proporcionar evidencia y visibilizar problemas que de otra forma no serían detectados, y garantizar el correcto funcionamiento de los sistemas. No obstante, una auditoría mal realizada puede facilitar la legitimación de sistemas riesgosos para los derechos. En los últimos años, se ha documentado que el diseño e implementación de auditorías suele ser cooptado mediante el '*audit-washing*'. Esto ocurre cuando las empresas o los gobiernos realizan evaluaciones superficiales que no permiten entender cómo funcionan los sistemas de manera significativa. Esto legitima prácticas controvertidas al instrumentalizar el principio de transparencia como un sello de confianza.

» **La necesidad de lograr una rendición de cuentas efectiva.** "Si la transparencia no tiene efectos significativos, la idea de transparencia pierde su propósito inicial"[75]. Anamy y Crawford consideran que la visibilidad puede poner en riesgo la rendición de cuentas si no existe un sistema que pueda procesar y utilizar esta información para promover un cambio. Tener toda la información disponible no implica que los actores estatales o privados tengan la intención de cambiar las políticas problemáticas de sus

[74] Artículo 40 del Reglamento de Servicios Digitales de la Unión Europea.

[75] Ananny, Mike, y Kate Crawford. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society 20*, no. 3 (marzo 2018): 973–89. https://doi.org/10.1177/1461444816676645.

sistemas automatizados. Muchas auditorías externas pueden realizar un diagnóstico que identifique todos los riesgos de un sistema automatizado, pero la rendición de cuentas resulta limitada si no existe un mecanismo que haga efectivas las demandas resultantes de esa detección de riesgos.

La transparencia es esencial pero insuficiente para garantizar sistemas automatizados predictivos justos, por lo que Debe reconcebirse no como un remedio único, sino como parte de un marco más amplio de responsabilidad estructural. Por ello debe servir a diferentes audiencias de distintas maneras: los reguladores necesitan documentación técnica; las personas afectadas requieren interfaces explicativas y mecanismos de impugnación; y la sociedad civil necesita acceso a herramientas de auditoría efectiva.

El choque de la transparencia algorítmica y el secreto industrial

La tensión entre los derechos de propiedad intelectual y el principio de transparencia algorítmica es común en las discusiones sobre desarrollo de sistemas automatizados respetuosos de los derechos humanos. Las tensiones se manifiestan especialmente en el *software* propietario, que cuenta con derechos de autor y cuya información sobre su desarrollo e implementación está frecuentemente protegida por el secreto industrial.

Como mencionamos anteriormente, tanto los agentes públicos como los privados a menudo invocan leyes de secreto industrial o comercial[76] para evitar cumplir con su deber de transparencia, pero estos principios deberían sopesarse siempre con el derecho del público a entender cómo se toman las decisiones que afectan de manera significativa a la ciudadanía.

En México, el secreto industrial protege la información confidencial con valor comercial o industrial que tenga una persona titular, siempre que brinde una ventaja competitiva y se

[76] Katherine J. Strandburg, Ryan Calo & Tal Zarsky, *Secret Algorithms, IP Rights, and the Public Interest, 21 Nev. L.J.* 61 (2020), https://scholars.law.unlv.edu/nlj/vol21/iss1/3/.

resguarde adecuadamente[77]. Se excluye la información de dominio público, de fácil acceso en el sector o que deba divulgarse por mandato legal. Sin embargo, entregar la información a una autoridad para obtener licencias o permisos no elimina su carácter confidencial si se mantiene protegida.

Respecto al uso de *software* propietario, el *código fuente* del algoritmo es uno de los elementos más sensibles desde la perspectiva de la propiedad intelectual, ya que consiste en el conjunto de instrucciones escritas en un lenguaje de programación legible por humanos[78]. Este código permite comprender, analizar, modificar y, eventualmente, replicar el funcionamiento del *software*, lo que podría otorgar una ventaja competitiva indebida a terceros.

En cambio, el *código objeto* es una versión compilada del código fuente, expresada en lenguaje de máquina y no legible por humanos. Este código constituye una etapa intermedia que, al ser enlazada, da lugar al código ejecutable o binario, que es la versión del *software* que normalmente se distribuye a los usuarios o terceros sin exponer su lógica interna[79]. Tanto el código fuente como el código objeto pueden estar protegidos por secreto industrial y, en algunos casos, esta protección también se extiende a su documentación técnica y a sus manuales de uso.

Sin embargo, en los casos en los que se utilice *software* propietario para la prestación de servicios públicos, el principio de máxima publicidad y el de rendición de cuentas deben prevalecer, sin que esto necesariamente se convierta en un juego de suma cero. Una solución a esta tensión podría ser habilitar licencias que incluyan auditorías externas controladas. Las autoridades podrían requerir en los contratos de licencia de uso de estos sistemas predictivos que se autorice a auditores externos, quienes estarían sujetos a las mismas responsabilidades de secreto industrial, pero que podrían divulgar los resultados de su análisis para evaluar el nivel de precisión, sesgos o información de interés público relacionada con los sistemas predictivos.

[77] Artículo 163 de la Ley Federal de Propiedad Industrial

[78] Carolina A. Canitrot, "Buenas Prácticas para la Protección de Software," Revista Iberoamericana de la Propiedad Intelectual 21 (2024): 354–355.

[79] Canitrot, Buenas prácticas, 354.

Para cumplir con sus obligaciones de transparencia social, los gobiernos deben elaborar informes de transparencia que incluyan la documentación necesaria para cumplir con los indicadores requeridos. Esta información puede incluir datos generales que a primera vista no comprometen el secreto industrial, tales como los nombres de los proveedores, las tasas de error o los resultados de evaluaciones de impacto o las auditorías. De acuerdo con el artículo 163, fracción primera de la Ley Federal de Protección a la Propiedad Industrial (LPPI), sería absurdo que las autoridades reservaran incluso el nombre del *Software* o la compañía que están utilizando, cuando esa misma información es de dominio público o generalmente conocida.

III. SISTEMAS PREDICTIVOS EN MÉXICO Y EL CAMINO HACIA LA TRANSPARENCIA ALGORÍTMICA

Transparencia Algorítmica en el Marco Jurídico Mexicano

El sistema de transparencia y protección de datos personales ha cambiado de manera drástica en el último año. Conceptos como la transparencia algorítmica se han manifestado de manera incipiente en distintas legislaciones, tales como las relacionadas con los medios alternativos de solución de controversias y las reformas para los trabajadores de plataformas.

Leyes de protección de datos personales y sus recientes reformas

En el primer trimestre del 2025, se aprobó un nuevo paquete de reformas en materia de protección de datos personales y transparencia. Ante la existente demanda de actualizar el marco jurídico desde hace décadas, estas reformas han desperdiciado la oportunidad de garantizar de manera efectiva la privacidad de la ciudadanía. Tanto la Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados (en adelante, Ley General) como la Ley Federal de Protección de Datos Personales en Posesión de Particulares (en adelante, Ley Federal) redujeron los estándares de protección a la privacidad, así como los mecanismos e instituciones que garantizaban dicha protección. Estas son las consideraciones relevantes del marco jurídico en materia de datos personales relacionadas con la transparencia algorítmica:

- **Aplicabilidad y rendición de cuentas.** Las nuevas reformas limitaron quiénes serán considerados sujetos obligados por la Ley General. Es decir, ya no se puede exigir la misma responsabilidad a un privado que realiza tratamiento automatizado con recursos

públicos que a una autoridad. Básicamente, se limita la posibilidad de que estos actores rindan cuentas mediante el juicio de amparo.

» **Consentimiento.** El principio de consentimiento y transparencia algorítmica están relacionados porque las personas deben estar debidamente informadas sobre quién, cómo y cuándo se tratarán sus datos personales. No obstante, la Ley General reduce los requisitos constitucionales para que un consentimiento sea realmente libre e informado.

Por un lado, la Ley General dificulta que las personas entiendan claramente qué datos se están recopilando. La nueva legislación disminuye los requisitos mínimos[80] del aviso de privacidad simplificado, que no cuenta con la información básica que cualquier persona usuaria necesita saber previo a consentir en el tratamiento de sus datos por las autoridades estatales.

Por otro lado, la Ley Federal ahora establece que las personas titulares dan su consentimiento de manera tácita cuando se pone a su disposición el aviso de privacidad". Para el caso de los datos sensibles, la ley sigue requiriendo el consentimiento expreso.

» **Oposición al tratamiento automatizado.** La Ley General mantiene el reconocimiento del derecho de las personas titulares a oponerse a tratamiento automatizado de datos personales[81]. Una persona puede solicitar acceder a la información[82] relacionada con el tratamiento automatizado de sus datos personales y decidir oponerse a estas prácticas. La oposición puede ejercerse si el tratamiento, aunque sea lícito, causa daño o perjuicio a las personas titulares, así como si el tratamiento automatizado genera efectos no deseados.

[80] Artículos 20 y 22 de la Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados.

[81] Artículo 41 de la LGPPDPPSO

[82] Artículo 38. La persona titular tendrá derecho de acceder a sus datos personales que obren en posesión del responsable, así como conocer la información relacionada con las condiciones y generalidades de su tratamiento.

» **Evaluación de Impacto en la Privacidad.** La Ley General en materia de Protección de Datos Personales contempla todavía la realización de una evaluación de impacto en la protección de datos personales[83] cuando se realice un tratamiento masivo de los mismos. Sin embargo, las recomendaciones que realice la Secretaría responsable no serán vinculantes.

Las evaluaciones de impacto en la protección de datos personales han sido una herramienta subutilizada por las autoridades mexicanas. Esto a pesar de que su importancia fue reconocida por la Suprema Corte de Justicia de la Nación en la Acción de Inconstitucionalidad 82/2021 respecto al Padrón Nacional de Usuarios de Telefonía Móvil[84]. La Corte resolvió que la ausencia de estas evaluaciones de privacidad expone los derechos a la privacidad, la intimidad y la protección de datos personales a un riesgo que no puede ser avalado a la luz de los artículos 6 y 16 de la Constitución[85]. Aunque estas evaluaciones son esenciales para la transparencia algorítmica, aún no abordan las diferentes afectaciones a derechos distintos del de la privacidad.

» **Falta de un organismo independiente.** La transparencia efectiva depende de una autoridad de protección de datos personales y transparencia que haga cumplir las obligaciones de las autoridades. El desmantelamiento del INAI representa un severo retroceso para alcanzar este objetivo.

[83] Fracción XV, Artículo 3, Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados.

[84] Suprema Corte de Justicia de la Nación (México). Acción de Inconstitucionalidad 82/2021 y su acumulada 86/2021. México: Suprema Corte de Justicia de la Nación, 2021. Disponible en: *https://www2.scjn.gob.mx/juridica/engroses/cerrados/Publico/Proyecto/AI82_2021y86_2021acumuladaPL.pdf*

[85] Suprema Corte de Justicia de la Nación, Acción de Inconstitucionalidad 82/2021, 159.

Reformas a la legislación en materia de transparencia y acceso a la información

La Ley General de Transparencia y Acceso a la Información Pública (Ley de Transparencia) eliminó varias garantías de transparencia que podrían contribuir a la transparencia algorítmica. Dentro de las más relevantes se encuentran:

- **Un sistema de transparencia en el que las excepciones a la transparencia son la regla.** Anteriormente, el artículo 11 establecía un estándar similar al del SIDH. Reconocía que el acceso a la información era la regla y las excepciones debían estar "definidas, legítimas y estrictamente necesarias en una sociedad democrática[86]. Este criterio interpretativo fue eliminado en la nueva Ley de Transparencia.

- **Debilitamiento de las obligaciones de documentación.** La Ley de Transparencia derogada requería que las autoridades generaran o repusieran la información que estaban obligadas a tener[87]. Ahora, la ley solo exige que se explique por qué no se tiene esa información y que se notifique a la persona solicitante[88].

  La transparencia algorítmica requiere que las autoridades generen informes y documentación de manera constante. A veces, a solicitud de auditores independientes o de la sociedad civil.

[86] Artículo 11 de la Ley General de Transparencia y Acceso a la información pública. Abrogada el 20 de marzo de 2025.

[87] Artículo 138, fracción III de la Ley General de Transparencia y Acceso a la información pública. Abrogada el 20 de marzo de 2025. "Ordenará, siempre que sea materialmente posible, que se genere o se reponga la información, en caso de que ésta tuviera que existir en la medida que deriva del ejercicio de sus facultades, competencias o funciones, o que previa acreditación de la imposibilidad de su generación, exponga de forma fundada y motivada, las razones por las cuales en el caso particular no ejerció dichas facultades, competencias o funciones (…)."

[88] Artículo 140, Fracción III la Nueva Ley General de Transparencia y Acceso a la información. Publicada el 20 de marzo de 2025.

La nueva ley todavía contempla los principios de máxima publicidad y documentación[89], con los que las personas pueden solicitar explicaciones de sistemas automatizados. No obstante, estos principios señalan que no se pueden requerir "documentos *ad hoc*".

» **Excepciones vagas y sobreinclusivas.** La reforma amplía el catálogo de información que puede clasificarse como reservada. Estas excepciones incluyen conceptos ambiguos como información que altere "la paz social[90]". Igualmente, se puede reservar la información sobre estudios o proyectos sin especificar o explicar de qué tipo de proyectos se trata, en casos donde ello pueda causar daños al interés del Estado o representar un riesgo para su ejecución[91]. Esta excepción es tan amplia que, en teoría, podría abarcar cualquier tipo de actividad realizada por un servidor público. La excepción establece un régimen de secrecía inicialmente y de transparencia posteriormente. En temas de transparencia algorítmica, esta fracción dificulta que se realicen evaluaciones previas o auditorías independientes que detengan el desarrollo de un sistema predictivo para poder atender sus riesgos o impactos frente a los derechos humanos.

Reformas relacionadas con los trabajadores de plataformas digitales

Las reformas a la Ley Federal de Trabajo relacionadas con los trabajadores de plataformas digitales establecen el principio de transparencia algorítmica. En el artículo 291-J se establece que las reglas para la asignación de servicios o tareas mediante algoritmos deben ser transparentes, claras y conocidas por los trabajadores. De igual forma, establece un mecanismo de transparencia social proactiva que exige que las plataformas elaboren un documento de política de gestión algorítmica del trabajo. Esta política debe explicar de manera clara y accesible

[89] Artículo 8, fracciones III y X de la Nueva Ley General de Transparencia y Acceso a la información. Publicada el 20 de Marzo de 2025.

[90] Artículo 112, Fracción I de la Nueva Ley General de Transparencia y Acceso a la información. Publicada el 20 de Marzo de 2025.

[91] Artículo 112, Fracción I de la Nueva Ley General de Transparencia y Acceso a la información. Publicada el 20 de Marzo de 2025.

las consecuencias de seguir instrucciones, el impacto de las calificaciones de terceros, los incentivos y penalizaciones, la existencia de categorías que influyen en la asignación de tareas y otros criterios relevantes. Este documento deberá integrarse al contrato laboral y ser conocido desde el inicio de la relación o ante cualquier modificación. Además, se establece que los algoritmos deben ser razonables, no poner en riesgo la salud o integridad del trabajador, ni generar discriminación.

De igual forma, en el artículo 291-P, se establece un mecanismo de apelación y rendición de cuentas para los trabajadores que debe ser gestionado por una persona y no por un sistema automatizado.

Transparencia Algorítmica en mecanismos alternativos de solución de controversias

El día 26 de enero de 2024, se publicó la nueva Ley General de Mecanismos Alternativos de Solución de Controversias en el Diario Oficial de la Federación[92]. Esta ley regula, en general, todas las formas de resolución de conflictos jurídicos, excepto los litigios, como el arbitraje o el proceso de conciliación. Dentro de esta ley se contempla un capítulo sobre la solución de controversias en línea que incluye las definiciones de sistemas automatizados y transparencia algorítmica.

El término 'sistemas automatizados'[93] en la ley funciona como un concepto general que abarca todo tipo de "inteligencia artificial". Esto incluye a sistemas de aprendizaje automático, cualquier tipo de sistema que realice procesamiento de datos, procesamiento de lenguaje natural, algoritmos y redes neuronales artificiales.

[92] Diario Oficial de la Federación. DECRETO por el que se expide la Ley General de Mecanismos Alternativos de Solución de Controversias y se reforma y adiciona la Ley Orgánica del Poder Judicial de la Federación y la Ley Orgánica del Tribunal Federal de Justicia Administrativa. Disponible en: *https://www.dof.gob.mx/nota_detalle.php?codigo=5715307&fecha=26/01/2024#gsc.tab=0*.

[93] Artículo 87, Fracción III de la Ley General de Mecanismos Alternativos de Solución de Controversias.

Igualmente, se hace un primer acercamiento al concepto de transparencia algorítmica. La legislación define este concepto como el conjunto de prácticas que hacen que los algoritmos utilizados por estos sistemas automatizados sean visibles, comprensibles y auditables. La legislación no define estos conceptos, por lo que no existe un criterio claro para aplicarlos. Adicionalmente, la transparencia algorítmica debe estar respaldada por una autoridad independiente que pueda evaluar y supervisar el uso de estos sistemas, lo cual sería un buen comienzo para mejorar la regulación en el futuro.

Un posible riesgo de esta implementación es que puedan existir sistemas automatizados en la toma de decisiones que formen parte de estos procesos. Esto puede afectar la igualdad entre las partes en ellos, aunque en principio el nivel de daño parecería menor, ya que ambas partes deben aceptar las condiciones de estos sistemas para poder continuar con ese MASC.

Transparencia Algorítmica en procesos jurisdiccionales

En agosto de 2025, el Segundo Tribunal Colegiado en Materia Civil del Segundo Circuito dictó la sentencia relativa a la queja civil 212/2025, en la cual se emplearon diversos sistemas generativos y se establecieron lineamientos para la utilización de inteligencia artificial en el ámbito jurisdiccional[94]. Aunque este informe se centra en modelos predictivos de inteligencia artificial, este caso es importante para comprender el marco legal de la transparencia algorítmica.

En este asunto, se buscaba calcular el monto de la garantía de acuerdo con los daños y perjuicios previstos en la jurisprudencia en la materia. El tribunal colegiado decidió calcular la garantía utilizando sistemas generativos comerciales, como *ChatGPT*, *Grok* y *Gemini*. La sentencia retomó los principios éticos de la inteligencia artificial de la UNESCO y algunos estándares de la Ley de IA europea.

[94] Segundo Tribunal Colegiado en Materia Civil del Segundo Circuito (México). *Sentencia relativa a la queja civil 212/2025*. México: Poder Judicial de la Federación, 29 de julio de 2025. Disponible en: *https://sise.cjf.gob.mx/SVP/word1.aspx?arch=103/01030000389400920002.pdf_1&sec=Juan_Carlos_Guerra_Alvarez&svp=1*

Sobre el principio de transparencia algorítmica, la sentencia estableció que toda persona juzgadora que utilice este tipo de sistemas en procesos jurisdiccionales debe informar sobre su uso. Además, se requiere que especifique los fines, los objetivos y los resultados alcanzados. La sentencia también incorporó el principio de supervisión humana y estableció que el uso de sistemas generativos no debe reemplazar la decisión de la persona juzgadora.

En la sentencia se justificó el uso de sistemas generativos y se describió la metodología empleada. El colegiado argumentó que estas herramientas se usaron a nivel operativo y no sustituyeron la labor jurisdiccional". De acuerdo con la sentencia, el uso de sistemas generativos para este caso permitió:

- La reducción de errores humanos en cálculos que no son en estricto sentido propios del razonamiento jurisdiccional.

- La transparencia y trazabilidad, al exponer el procedimiento que se siguió para llegar al resultado numérico.

- La estandarización y coherencia en los precedentes utilizados.

- Un mayor grado de eficiencia procesal.

Posteriormente, la sentencia incluyó el *prompt* que se ingresó en los diferentes sistemas generativos comerciales y comparó los resultados. El prompt consiste en la fórmula establecida por los criterios jurisprudenciales para realizar el cálculo. Esta fórmula no tiene valores fijos, ya que deben obtenerse del expediente en concreto y de páginas oficiales como el INEGI y el Banco de México. Las cantidades coinciden en los tres sistemas y la sentencia retoma estas cantidades para fijar la garantía.

Estas instrucciones parecen tener como objetivo crear una automatización para un sistema de cálculo. A primera vista, el uso de un sistema automatizado para el cálculo de una garantía no parece un uso intensivo o desproporcionado. La sentencia indica que las instrucciones y las respuestas incluyen una tabla explicativa y auditable que muestra, paso a paso, cómo se obtuvo el resultado. Por lo tanto, de acuerdo con la sentencia, se preserva el núcleo esencial de la función

jurisdiccional. No obstante, la sentencia no especifica cuál es el núcleo esencial de esta función ni los parámetros para analizarlo, aunque podemos deducir que se refiere a que no es un factor determinante que sustituya la decisión del juzgador.

Desde un enfoque de derechos, este ejercicio genera dudas sobre si se cumplieron los requisitos básicos de las obligaciones de transparencia. Particularmente, las personas que toman decisiones siempre deben consultar a las partes involucradas para saber si consienten en el uso de los sistemas generativos, en su caso. Los sistemas comerciales generativos pueden cometer errores o, en ocasiones, requieren información sensible para obtener estos resultados. La falta de notificación limita la capacidad de las personas justiciables para alegar su inconformidad y violaría entonces una de las formalidades esenciales del proceso.

También llama la atención los límites de este tipo de transparencia. Parece que se refiere a la transparencia relacionada con la persona que la utiliza, pero no necesariamente a la transparencia de la herramienta en sí. Aunque podemos analizar el *prompt* y las respuestas, una persona no puede realizar una evaluación técnica del proceso algorítmico. Por ejemplo, no podemos saber si el código en efecto accedió a las páginas para obtener esa información, ni saber si la información usada es veraz y no fue una *confabulación*[95] que resultó en información incorrecta.

El análisis de todas las responsabilidades de la persona juzgadora al usar sistemas generativos puede ser objeto de una futura línea de investigación. No obstante, es interesante observar cómo estos principios van permeando en el Poder Judicial y los desafíos que se avecinan en futuras regulaciones.

[95] Otro término para referirse a la confabulación es "alucinación". No obstante, este término ha sido criticado por antropomorfizar a los sistemas automatizados. Para más información, consultar: "Why AI Chatbots Are the Ultimate BS Machines — and How People Hope to Fix Them," *Ars Technica*, abril de 2023. Disponible en: *https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/*

Estudio de casos de sistemas predictivos en México

SESNA: Algoritmo para el uso de servicios sociales

En el año 2024, un grupo de investigadores del CIDE solicitó información a través de transparencia a distintas entidades gubernamentales, sobre el uso de sistemas automatizados en el ejercicio de sus funciones. El resultado de esta investigación se consolidó en el repositorio de Algoritmos Públicos[96].

En respuesta a una solicitud de acceso a la información[97], la Secretaría Ejecutiva del Sistema Nacional Anticorrupción (SESNA) informó que estaba desarrollando un proyecto relacionado con algoritmos en programas sociales. SESNA establece diversas prioridades dentro de la Política Nacional Anticorrupción (PNA), entre las cuales se incluye promover el uso de tecnologías avanzadas, como el análisis de datos masivos y la inteligencia artificial, con el fin de identificar riesgos y mejorar la gestión, auditoría y fiscalización estratégica en el sector público.

Asimismo, la PNA propone fortalecer la evaluación de los programas presupuestarios incorporando enfoques de derechos humanos y mecanismos para gestionar los riesgos de corrupción. También plantea la creación de observatorios y laboratorios de innovación social enfocados en detectar y prevenir riesgos de corrupción, especialmente en los puntos de contacto entre el gobierno y la sociedad, así como en los procesos de compras y adquisiciones públicas.

Dentro de la respuesta a la solicitud de información, SESNA no clasificó este sistema automatizado como IA, aunque no parece ser un sistema de aprendizaje automático en sentido estricto. Por ejemplo, no hay una mención explícita de entrenamiento de modelos con datos etiquetados, pero sí se utilizan principios y herramientas comunes en este campo, tales como

[96] Algoritmos CIDE, *Repositorio - Algoritmos* CIDE, Centro de Investigación y Docencia Económicas (CIDE), última consulta 2 de julio de 2025, https://algoritmoscide.org/repositorio/.

[97] Respuesta a la solicitud de acceso a la información 331637024000047, Oficio SE/UT/0078/2024 de la Secretaría Ejecutiva del Sistema Nacional Anticorrupción. 15 de abril de 2024. Este documento se encuentra en el repositorio de Algoritmos CIDE, disponible en https://osf.io/cu6w3/files/osfstorage

simulación basada en datos, evaluación por proximidad y análisis de agrupamientos. De cualquier manera, sigue siendo un sistema algorítmico con componentes inspirados en el aprendizaje automático o que puede escalarse a un sistema mucho más complejo en el futuro.

Respecto al interés público que puede generar este algoritmo, es importante señalar que el algoritmo que utilizan produce simulaciones que clasifican a las personas según su nivel de vulnerabilidad, y podría ser utilizado para tomar decisiones sobre la eficiencia en la distribución de servicios públicos. Aunque parece que no es un algoritmo completamente determinista, es necesario que expliquen cómo se utiliza.

El proyecto aún está en desarrollo, por lo que no hay más información técnica disponible. Si aplicáramos los estándares de transparencia, en esta etapa el SESNA debería estar en consulta con diversas partes interesadas para analizar los criterios con los que se busca entrenar este sistema y alcanzar los objetivos necesarios.

Sistemas predictivos establecidos en la nueva ley de inteligencia

El primero de julio de 2025, se creó la Ley del Sistema Nacional de Investigación e Inteligencia en Materia de Seguridad Pública. Esta reforma se aprueba en el contexto de una serie de iniciativas que crean una infraestructura de vigilancia masiva en México.

El paquete de reformas incluye la Ley General de Población y la Ley General en Materia de Desaparición Forzada de Personas, Desaparición Cometida por Particulares y del Sistema Nacional de Búsqueda de Personas. A partir de ellas, se establece la creación de una identificación biométrica obligatoria para toda la población, así como la creación de la Plataforma Nacional de Identidad. La nueva Cédula de Identidad Única registrará en tiempo real cada vez que se utilice para acceder a servicios públicos o privados.

Adicionalmente, en la Ley de Inteligencia se crea una Plataforma Central de Inteligencia[98] que se conectrá con diferentes bases de datos públicas y privadas, formando un único nodo

[98] Artículo 2, Fracción VII de la Ley del Sistema Nacional de Investigación e Inteligencia en Materia de Seguridad Pública.

de información. La recolección masiva y centralizada de datos personales resulta aún más preocupante debido a las facultades legales que la ley otorga a las autoridades de seguridad pública. El artículo 32 establece que las autoridades pueden utilizar herramientas de inteligencia artificial para procesar y analizar información recabada o utilizada por la plataforma con el fin de prevenir y perseguir delitos.

Los sistemas predictivos para la prevención del delito han sido constantemente señalados como herramientas sesgadas, falibles y discriminatorias[99]. Uno de los sistemas más conocidos es *COMPAS*. Este *software* afirmaba poder prevenir el riesgo de reincidencia o la posibilidad de fuga de las personas acusadas en un proceso penal.

En 2016, *ProPublica* publicó una investigación que demostró cómo este sistema era defectuoso y tenía un sesgo contra las personas afroamericanas[100]. De las personas que el sistema predijo que reincidirían en cometer delitos violentos, solo el 20 % realmente reincidió. El código etiquetaba incorrectamente a las personas afroamericanas, asignándoles el doble de probabilidad de cometer delitos en comparación con las personas blancas.

En general, los sistemas predictivos que evalúan el riesgo de reincidencia son considerados como incompatibles con los derechos humanos. A tal grado que la Ley de IA de la Unión Europea prohíbe su oferta o puesta en operación[101]. En este caso, nos encontramos en casos de usos no permitidos por el DIDH durante la fase de concepción. Por lo tanto, las autoridades mexicanas deberían abstenerse de utilizar estas herramientas.

[99] Renata M. O'Donnell, "Challenging Racist Predictive Policing Algorithms under the Equal Protection Clause," New York University Law Review 94, no. 3 (June 2019): 544-580

[100] Julia Angwin, Jeff Larson, Surya Mattu y Lauren Kirchner, "Machine Bias: Risk Assessments in Criminal Sentencing," *ProPublica*, 23 de mayo de 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[101] Unión Europea. *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de junio de 2024 relativo a normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y por el que se modifican determinados actos legislativos de la Unión (AI Act).* Diario Oficial de la Unión Europea, L 168/1, 12 de julio de 2024. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32024R1689.

Como mínimo, los gobiernos tienen que revelar de manera proactiva si consideran utilizar o utilizan este tipo de herramientas. Esta es una afectación grave a los derechos de no discriminación, debido proceso y presunción de inocencia de la ciudadanía. En caso de que ya esté en funcionamiento, las personas deben ser notificadas si son sometidas a estos sistemas predictivos. Este tipo de herramientas no tiene cabida en un país libre y democrático. Es urgente reformar aquellas disposiciones que permiten tanto la recolección masiva e indiscriminada de datos personales, como cualquier referencia que permita la creación y el uso de estos sistemas predictivos.

Sistemas predictivos en procesos de recaudación tributaria

En el año 2024, el Servicio de Administración Tributaria (SAT) anunció[102] el uso de inteligencia artificial para mejorar la planificación en los procesos de recaudación. El objetivo es implementar modelos de analítica de grafos y aprendizaje automático para clasificar a los contribuyentes según su riesgo, identificar redes complejas de elusión y evasión, así como detectar inconsistencias en los comprobantes fiscales digitales (CFDI).

Este sistema predictivo busca asignar un nivel de riesgo de incumplimiento a los contribuyentes utilizando criterios desconocidos. El SAT maneja una gran cantidad de datos de la ciudadanía: comprobantes fiscales digitales, declaraciones, información bancaria e incluso datos biométricos.

Pedro David Nieto considera que este tipo de información genera un riesgo importante de sesgos algorítmicos[103]:

- Los CFDI a menudo se emiten con errores humanos y pueden introducir datos incorrectos en los sistemas predictivos. Igualmente, excluyen gran parte de la economía informal, lo que puede provocar una subrepresentación de ciertos sectores.

[102] Servicio de Administración Tributaria, Plan Maestro 2024, diapositiva 9.

[103] Pedro David Nieto Olvera, "Artificial Intelligence and Algorithms in Tax Auditing by the Tax Administration Service in Mexico: Analysis of Potential Biases," *International Journal for Public Policy, Law and Development 2*, no. 3 (2025).

- Las declaraciones fiscales son realizadas por los contribuyentes, por lo que su confiabilidad puede variar y pueden reflejar desigualdades en el acceso a asesoría o en el nivel de alfabetización fiscal. De igual forma, no es información particularmente útil para predecir el futuro.

- El acceso a la información bancaria es particularmente sensible en materia de protección de datos personales; se trata del análisis automatizado de información sobre los aspectos más personales de nuestra vida. Un sistema predictivo entrenado con clasificaciones sesgadas puede penalizar injustamente a los contribuyentes por diferencias en su comportamiento financiero relacionadas con su situación socioeconómica, tipo de negocio o sector. Los datos de terceros, como los provenientes del IMSS, INFONAVIT o registros públicos, pueden ser inconsistentes o llegar con retraso, lo que genera errores que afectan al contribuyente auditado sin que este tenga responsabilidad directa.

El SAT ha sido poco transparente en el uso de estas herramientas. En el año 2021, un investigador reportó haber realizado varias solicitudes de acceso a la información para conocer cómo se utilizaba la IA dentro del SAT[104]. La dependencia rechazó tener este tipo de información.

Este tipo de sistemas predictivos no fue notificado directamente a la ciudadanía. Por lo tanto, desconocemos en qué momento se empezaron a utilizar los datos personales de la ciudadanía para entrenar a sus sistemas predictivos. A pesar de esto, el sistema predictivo está procesando la información de las personas y tomando decisiones que afectan sus derechos.

El SAT debe de desglosar esta información para cumplir con sus obligaciones en materia de transparencia algorítmica:

- Explicar en detalle qué tan determinante es el uso de estas herramientas para decidir quién será sometido a una auditoría.

[104] Eugenio Argüelles Toache, "Beneficios y riesgos del uso de la Inteligencia Artificial en el Servicio de Administración Tributaria de México (SAT). Un análisis desde la perspectiva de investigadores académicos," PAAKAT: *Revista de Tecnología y Sociedad* 14, no. 27 (septiembre 2024), publicado electrónicamente el 22 de octubre de 2024, https://doi.org/10.32870/pk.a14n27.885.

- Abrir espacios para que auditores independientes puedan evaluar el funcionamiento de esta herramienta. Así como dar a conocer cualquier evaluación de impacto en la privacidad realizada, especialmente porque es una obligación legal que debieron haber cumplido antes de su operación.

- Indicar si está utilizando estos sistemas predictivos para evaluar el riesgo fiscal de los contribuyentes.

- Notificar a los contribuyentes de manera accesible e individual si su información fue procesada por estos sistemas predictivos. Asimismo, debe notificarles en caso de que se haya tomado una decisión mediante estos sistemas.

- Explicar claramente a las personas afectadas por estos sistemas sobre los factores que influyeron en la decisión. En caso de que la persona desee apelar la decisión, debe tener la oportunidad de analizar el sistema predictivo de manera independiente.

RECOMENDACIONES FINALES

A lo largo de este informe, describimos las distintas obligaciones en materia de transparencia y acceso a la información para el uso de sistemas automatizados en la toma de decisiones. Por lo tanto, creemos que son necesarias, como mínimo, las siguientes acciones para mejorar significativamente la transparencia algorítmica en México.

1 Tomarse en serio las obligaciones relacionadas con la transparencia y los derechos humanos

La principal recomendación es que las autoridades estatales sean responsables de sus obligaciones legales en materia de derechos humanos. A partir de ellas eben dejar de plantear sus obligaciones como criterios éticos o voluntarios, para dotarlas de una eficacia legal plena que permita establecer responsabilidades de distinto tipo a los entes públicos y a los sujetos privados con relación al ámbito público.

Existen ciertos sistemas automatizados que, por su naturaleza, no son compatibles con los derechos humanos. En principio, los gobiernos deben evitar implementar este tipo de tecnología a toda costa. Asimismo, las reservas absolutas de cualquier tipo de información relacionada con el uso de sistemas predictivos no solo son contrarias a los derechos humanos, sino también a los principios democráticos fundamentales, como la rendición de cuentas y la transparencia en el servicio público.

El Estado Mexicano debe justificar el uso de sistemas predictivos para cumplir con sus responsabilidades antes de su implementación, y notificar en aquellos casos en que decida emplearlos. Esta justificación debe ajustarse a los estándares de derechos humanos, implicando que:

- Las autoridades deben reunirse con las distintas partes interesadas para discutir la legalidad, necesidad y proporcionalidad del sistema predictivo.

- En caso de que el objetivo final cumpla con estos estándares, su uso deberá regularse de manera clara y precisa en la legislación.

- Se debe notificar a las personas usuarias antes de la etapa de implementación del sistema automatizado. Es importante que las personas puedan consentir en el uso de sus datos personales.

- Realizar una evaluación integral del impacto en los derechos humanos. Este análisis debe realizarse antes de cualquier avance en el desarrollo del sistema predictivo y en colaboración con las múltiples partes interesadas.

2 La transparencia debe ser la norma, por lo que debe establecerse un esquema de transparencia activa y efectiva cuando se utilizan sistemas predictivos en la toma de decisiones

El Estado debe implementar políticas públicas que se traduzcan en medidas legales y técnicas para que la transparencia sea la norma.

Medidas legales:

- Adaptar la legislación para evitar la opacidad en la información relacionada con sistemas predictivos. La transparencia efectiva implica que el gobierno no debe reservar información de manera absoluta. Esto implica que deben reformarse los artículos que permitan negar información de manera categórica bajo el argumento de seguridad nacional. Ejemplos como este los vemos en las nuevas reformas a las leyes de seguridad e inteligencia, donde la reserva total de información por cuestiones de seguridad nacional pone en riesgo cualquier tipo de transparencia algorítmica relacionada con el uso de sistemas predictivos para la prevención del delito.

- Evitar caer en el falso dilema de cumplir únicamente con sus obligaciones en materia de transparencia o con las de propiedad intelectual. No se puede hacer una interpretación estricta de figuras legales como el secreto industrial. En este sentido, las autoridades también deben asegurarse de que los contratos de licencia de *software* cuenten con flexibilidades para que auditores externos puedan realizar sus funciones de supervisión.

Medidas técnicas:

- Evitar el uso de sistemas predictivos con algoritmos que no sean interpretables o que dificulten la rendición de cuentas.

- Implementar mecanismos de supervisión mediante evaluaciones de impacto previas y auditorías en las diferentes etapas del ciclo de vida de la IA. Las auditorías deben realizarse con auditores independientes que incluyan a las distintas partes interesadas.

- Asegurarse de documentar de manera activa la información relacionada con los sistemas predictivos. Es decir, no deben esperar a que las partes interesadas o auditores les soliciten generar información específica. De igual forma, deben generarse informes periódicos sobre las distintas actividades o resultados relacionados con el uso de sistemas predictivos.

3 Evitar soluciones cosméticas que permitan a los actores estatales eludir su responsabilidad

Los actores políticos deben resistirse a la tentación de tomar soluciones superficiales. En su lugar, deben implementar salvaguardas sólidas que aseguren una rendición de cuentas efectiva. Esto incluye garantizar una transparencia significativa, así como los principios de participación, supervisión y reparación. Solo entonces podrán los sistemas algorítmicos de las infraestructuras públicas servir realmente a fines democráticos y respetuosos de los derechos humanos.

Deben tenerse claros los niveles de responsabilidad durante el desarrollo e implementación de los sistemas predictivos. Esto incluye tener claramente definidos quiénes son los puntos de contacto en cada etapa y cuáles son los distintos grados de responsabilidad que tienen los sujetos que participan en el desarrollo y uso de los sistemas. Por ejemplo, las personas que se dedican a clasificar y entrenar a los sistemas predictivos durante la fase de desarrollo no tienen la misma responsabilidad que una funcionaria que utilice esta herramienta en la fase de operación. Las auditorías o evaluaciones constantes de estos sistemas son útiles para identificar estas responsabilidades desde el principio.

4 La participación efectiva y abierta de diferentes actores es esencial para el éxito de cualquier programa que implemente sistemas predictivos en servicios públicos

Es importante que los gobiernos abran las puertas a la sociedad civil, personas expertas y, en especial, a las personas más afectadas por estos procesos. Esta transparencia social no solo evitará riesgos innecesarios para los derechos humanos, sino que también generará mayor confianza en la sociedad que convivirá día a día con estos sistemas automatizados.

TRANSPARENC*IA* ALGORÍTMICA

Obligaciones de Derechos Humanos en las Decisiones Automatizadas

Escrito por: Grecia Macías

Portada e interiores: Andrés Timm

Noviembre de 2025

R3D
Red en Defensa
de los Derechos Digitales

Brot
für die Welt