FORSETI

Forseti. Revista de Derecho. Volumen 14, Nº 22, Lima, 2025, pp. 227 – 245

Algoritmos opacos y privacidad diferencial: ¿puede haber transparencia sin comprometer la protección?

Juan Emmanuel Delva Benavides*
y Vania Valeria Mendez Gonzalez**

Resumen.- El presente artículo analiza la tensión entre transparencia algorítmica y privacidad diferencial en el contexto de implementación creciente de sistemas de Inteligencia Artificial en ámbitos sensibles. A través de un análisis interdisciplinario que integra perspectivas jurídicas, tecnológicas y éticas, se examina cómo los algoritmos opacos comprometen la rendición de cuentas y cómo la privacidad diferencial, aunque protege datos personales, puede socavar la exactitud algorítmica.

Palabras clave.- Algoritmos opacos – privacidad diferencial – transparencia algorítmica – protección de datos – inteligencia artificial – derecho digital.

Abstract.- This article analyzes the tension between algorithmic transparency and differential privacy in the context of increasing implementation of Artificial Intelligence systems in sensitive areas. Through an interdisciplinary analysis that integrates legal, technological, and ethical perspectives, it examines how opaque algorithms compromise accountability and how differential privacy, while protecting personal data, can undermine algorithmic accuracy.

Keywords.- Opaque algorithms – Differential privacy – Algorithmic transparency – Data protection – Artificial intelligence – Digital law.

* Profesor Investigador en la Universidad de Guadalajara. Doctor en Derecho, maestro en tecnologías y abogado por parte de la UdeG. Socio fundador de la Asociación Latinoamericana de Derecho, Gobierno y Nuevas Tecnologías.

** Estudiante de la carrera de Negocios Internacionales de la Universidad de Guadalajara y técnico en programación por parte del Centro de Bachilleres Industriales y Tecnológicos.

I. Introducción

Desde inicios del siglo XXI, la humanidad está experimentando una transformación digital sin precedentes, comparable quizás solo con la revolución industrial del siglo XVIII. La diferencia fundamental radica en que, mientras la revolución industrial transformó nuestra relación con el mundo físico, la revolución digital está reconfigurando nuestra existencia informacional y cognoscitiva. A modo de ilustración, si la máquina de vapor permitió multiplicar la fuerza física humana, los algoritmos de inteligencia artificial están amplificando —y en algunos casos, sustituyendo— nuestras capacidades cognitivas y decisorias.

La sociedad contemporánea se caracteriza por una hiperconectividad creciente[1], donde el trabajo, la educación, el comercio e incluso las interacciones sociales están mediados por tecnologías digitales. Esta transformación se refleja en cifras contundentes: 4.88 billones de usuarios de Internet a nivel global en 2021, con un 62% de la población mundial utilizando redes sociales[2]. En México, el panorama no es diferente: el INEGI (2024) reporta que el 68.5% de los hogares mexicanos —equivalente a 25.8 millones— cuenta con acceso a Internet [3].

Esta digitalización masiva ha catalizado el surgimiento de lo que podríamos denominar la "economía del dato", un ecosistema donde la información personal se ha convertido en el petróleo del siglo XXI. En este contexto, los corredores de datos o *data brokers* operan como nuevos intermediarios, recolectando y comercializando información de diversas fuentes para su posterior análisis. Al igual que ocurrió con los recursos naturales en siglos anteriores, estamos presenciando una "carrera por el dato" donde empresas y gobiernos compiten por capturar, procesar y monetizar la mayor cantidad posible de información personal.

---

1 DURAO, M.; ETCHEZAHAR, E.; GÓMEZ, T.; MULLER, M. (2024). *Hiperconectados: Los desafíos psicológicos de la era digital*. Madrid, España: TechPsyLab

2 DataReportal. (2021). Digital 2021 october global statshot report. https://datareportal.com/reports/digital-2021-october-global-statshot

3 INEGI (Instituto Nacional de Estadística y Geografía). (2024). Estadísticas a propósito del Día Mundial del Internet. https://inegi.org.mx/contenidos/saladeprensa/aproposito/2024/EAP_DMInternet.pdf

El aprovechamiento comercial de estos datos ocurre a través de múltiples vías: desde políticas de privacidad que autorizan la cesión a terceros, hasta tecnologías como las *cookies* que monitorizan en tiempo real los patrones de comportamiento digital. Como en el caso del legendario caballo de Troya, detrás de la aparente gratuidad de muchos servicios digitales se esconde un mecanismo sofisticado de captura de datos personales.

Las empresas suelen justificar estas prácticas argumentando que los datos son "anonimizados", eliminando identificadores directos. Sin embargo, como demuestra Narayanan[4], existen numerosos "vectores de fuga" que posibilitan la reidentificación de individuos mediante la correlación de diferentes conjuntos de datos aparentemente anónimos. Es comparable a un rompecabezas donde, aunque faltan piezas individuales, la imagen completa puede reconstruirse mediante la combinación inteligente de fragmentos dispersos.

Es en este contexto donde emergen y adquieren relevancia crítica los conceptos de algoritmos opacos y privacidad diferencial, representando dos valores aparentemente contrapuestos: por un lado, la necesidad de transparencia en los sistemas algorítmicos que toman decisiones con impacto significativo en la vida de las personas; por otro, la imprescindible protección de la privacidad en un entorno de datificación masiva.

El presente artículo examina esta tensión fundamental y busca responder a una pregunta central para el Estado de Derecho en la era digital: ¿Es posible construir sistemas algorítmicos que sean simultáneamente transparentes —y, por tanto, auditables— y respetuosos de la privacidad? ¿O estamos ante un dilema irresoluble donde debemos sacrificar un valor en aras del otro?

La tensión entre algoritmos opacos y privacidad diferencial representa, más que un dilema técnico pasajero, una manifestación concreta del conflicto fundamental que toda sociedad democrática debe resolver en la era digital: cómo equilibrar transparencia, privacidad, seguridad y eficiencia. Esta disyuntiva evoca la clásica tensión entre libertad

[4] NARAYANAN, A. (2011). There is no such thing as anonymous online tracking. The Center for Internet and Society at Stanford Law School. http://cyberlaw.stanford.edu/blog/2011/07/there-no-such-thing-anonymous-online-tracking. (Fecha de consulta: 02 de mayo de 2025).

y seguridad que ha definido los debates constitucionales desde la Ilustración, pero con un elemento distintivo y perturbador: la velocidad vertiginosa con que la tecnología reconfigura los términos del debate, comparable quizás a cómo la imprenta de Gutenberg transformó las estructuras de poder informacional en el siglo XV, pero a una escala exponencialmente acelerada. Mientras que la sociedad tardó décadas en adaptar sus estructuras regulatorias a la imprenta, hoy enfrentamos cambios paradigmáticos en ciclos de meses o incluso semanas.

Dicho esto, este trabajo parte de la hipótesis de que a pesar de la tensión entre la privacidad de los usuarios y la necesidad de la transparencia en el tratamiento de dichos datos -para que sean auditables-, es posible diseñar mecanismos que permitan mantener ambas variables sin que una comprometa a la otra.

II. Marco Teórico Conceptual: Entre la Opacidad y la Distorsión

i. Algoritmos opacos: Las cajas negras decisorias

Los algoritmos opacos representan una de las paradojas más significativas de la era digital: sistemas cada vez más determinantes en nuestras vidas, pero simultáneamente inescrutables en su funcionamiento interno. El término "caja negra", prestado de la ingeniería de sistemas, ilustra perfectamente esta contradicción: podemos observar lo que entra (datos) y lo que sale (decisiones), pero el proceso intermedio permanece oculto, incluso para sus propios creadores.

Esta opacidad no es meramente técnica, sino inherente a la naturaleza de los sistemas de aprendizaje automático contemporáneos (*Machine Learning*). A diferencia de la programación tradicional, donde un desarrollador codifica explícitamente cada regla lógica, el aprendizaje automático genera sus propios modelos a partir del análisis de enormes cantidades de datos, creando redes neuronales complejas cuyo funcionamiento interno resulta prácticamente indescifrable[5].

Para dimensionar adecuadamente el problema, imaginemos un juez humano que tomara decisiones sin explicar jamás su razonamiento, o un médico que diagnosticara enfermedades sin revelar qué síntomas

[5] HURSHMAN, C. "¿Tienen funciones los algoritmos opacos?" En Synthese, 204, 91 (2024). https://doi.org/10.1007/s11229-024-04745-2

o pruebas considera relevantes. Tal escenario resultaría inaceptable en contextos profesionales humanos, pero es precisamente lo que ocurre con muchos sistemas algorítmicos que ya están operando en ámbitos críticos como la justicia penal, la medicina diagnóstica o la evaluación crediticia.

ii. Privacidad diferencial: El arte de ocultar mostrando

Como contrapunto a la transparencia algorítmica, la privacidad diferencial emerge como un enfoque matemático que busca proteger la información personal mientras permite el análisis agregado de datos. Desarrollada inicialmente por Dwork[6], esta metodología utiliza la distorsión controlada —añadiendo "ruido" estadístico calibrado— para garantizar que la presencia o ausencia de un individuo en una base de datos no afecte significativamente los resultados del análisis.

La privacidad diferencial podría compararse con una técnica fotográfica que deliberadamente desenfoca ciertos elementos de la imagen para proteger identidades específicas, manteniendo a la vez la nitidez suficiente para interpretar la escena general. El desafío radica en determinar el nivel óptimo de "desenfoque" que proteja adecuadamente sin comprometer excesivamente la utilidad analítica de los datos.

iii. La dialéctica entre transparencia y privacidad

La tensión entre transparencia algorítmica y privacidad diferencial puede conceptualizarse como una relación inversamente proporcional: cuanto mayor es la transparencia, mayor el riesgo para la privacidad; y cuanto más robusta la protección de privacidad, menos precisos o transparentes resultan los algoritmos.

Esta relación inversamente proporcional evoca el principio de incertidumbre de Heisenberg en física cuántica[7]; así como es imposible

---

[6] DWORK, C. (2006). Differential Privacy. 33rd International Colloquium on Automata, Languages and Programming, pp. 1-12.

[7] HEISENBERG, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. Zeitschrift für Physik, 43(3-4), pp. 172-198. Esta referencia alude al principio físico que establece la imposibilidad de determinar simultáneamente con precisión absoluta la posición y el momento de una partícula. La analogía aplicada al contexto de algoritmos sugiere que existe una limitación

conocer simultáneamente y con precisión absoluta la posición y el movimiento de una partícula, parece existir una limitación fundamental para lograr simultáneamente máxima transparencia algorítmica y máxima protección de la privacidad.

III. Tensión entre Transparencia y Privacidad: Un Análisis Multidimensional

i. El imperativo ético de la transparencia

La transparencia algorítmica constituye un requisito fundamental para la rendición de cuentas en sociedades democráticas. Cuando los algoritmos determinan quién obtiene un préstamo, quién recibe tratamiento médico prioritario o quién representa un riesgo para la seguridad pública, su funcionamiento no puede permanecer como una "caja negra" impenetrable.

El caso emblemático de COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) en Estados Unidos ilustra claramente esta problemática. Este sistema, utilizado para evaluar el riesgo de reincidencia criminal, ha sido objeto de intenso escrutinio tras el análisis de ProPublica[8] que reveló sesgos raciales significativos. Sin embargo, la empresa desarrolladora se ha negado a revelar completamente el funcionamiento interno del algoritmo, argumentando protección de propiedad intelectual y secretos comerciales.

Este caso evidencia cómo la opacidad algorítmica puede socavar principios fundamentales del debido proceso: ¿cómo puede un acusado defenderse efectivamente si desconoce la lógica que lo clasifica como "alto riesgo"? Es comparable a un juicio donde la prueba determinante permanece sellada, inaccesible tanto para el acusado como para su defensa.

---

fundamental entre la transparencia completa de un sistema algorítmico y la protección total de la privacidad de los datos que utiliza.

[8] ANGWIN, J; LARSON, J; MATTU, S. & KIRCHNER, L. (2024). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica.

ii. El imperativo legal de la privacidad

Por otra parte, la protección de datos personales constituye un derecho fundamental reconocido en marcos normativos como el Reglamento General de Protección de Datos (GDPR) europeo o la Ley Federal de Protección de Datos Personales en Posesión de Particulares en México. Estos marcos exigen que los datos personales sean tratados con medidas técnicas y organizativas que garanticen su confidencialidad e integridad.

La privacidad diferencial ofrece una solución técnica prometedora para este imperativo legal. Sin embargo, como demuestra Salaberry[9], esta metodología implica un compromiso inevitable entre privacidad y precisión: cuanto más se protege la privacidad mediante la adición de "ruido" estadístico, menos precisos resultan los datos para análisis posteriores.

iii. La paradoja de la explicabilidad computacional

La tensión entre transparencia y privacidad se manifiesta técnicamente en lo que podríamos denominar "la paradoja de la explicabilidad". Los investigadores Tramer[10] y Shokri et al.[11] han demostrado que los modelos completamente transparentes son vulnerables a lo que se conoce como "ataques de inversión" o "ataques de membresía", mediante los cuales un adversario puede inferir si determinados datos individuales formaron parte del conjunto de entrenamiento del modelo.

Este fenómeno es comparable a cómo un estudiante podría deducir el contenido exacto de un examen futuro si tuviera acceso completo al proceso mental del profesor que lo diseña. La transparencia total, paradójicamente, puede comprometer la integridad del sistema que se pretende explicar.

---

9 SALABERRY, N. (2020). Gestión de la privacidad de datos personales: el modelo de privacidad diferencial. Revista de Investigación En Modelos Matemáticos Aplicados A la Gestión y la Economía, II (2020-II), pp. 45-67.

10 TRAMER, F. (2023). Modelo de inversión de ataques: Desafíos para la privacidad en la era de modelos transparentes. IEEE Symposium on Security and Privacy, 432-449.

11 SHOKRI, R; STRONATI, M; SONG, C. & SHMATIKOV, V. (2021). Membership Inference Attacks Against Machine Learning Models. IEEE Symposium on Security and Privacy, pp. 3-18.

IV. Casos de estudio sectoriales

La tensión entre transparencia algorítmica y privacidad diferencial se manifiesta de manera particular en sectores críticos donde los algoritmos están transformando radicalmente los procesos decisorios. Al igual que la invención de la imprenta revolucionó el acceso al conocimiento en el siglo XV, los algoritmos de aprendizaje profundo están reconfigurando industrias enteras, pero con una diferencia fundamental: mientras que los libros impresos exponían abiertamente su contenido, los algoritmos contemporáneos ocultan sus mecanismos internos tras capas de complejidad matemática.

En el sector sanitario, asistimos a una revolución diagnóstica comparable a la introducción del microscopio en el siglo XVII. Los algoritmos de diagnóstico médico basados en redes neuronales convolucionales, como los desarrollados por Google Health para la detección de retinopatía diabética, han demostrado una precisión diagnóstica que iguala o supera a oftalmólogos experimentados[12]. Sin embargo, a diferencia del médico que puede explicar paso a paso su razonamiento clínico, estos sistemas funcionan como "oráculos digitales" que emiten veredictos sin revelar su lógica interna. Esta opacidad plantea dilemas éticos fundamentales: ¿cómo puede un paciente otorgar consentimiento verdaderamente informado cuando ni siquiera los desarrolladores del algoritmo comprenden completamente por qué el sistema llega a determinada conclusión? Simultáneamente, la extrema sensibilidad de los datos médicos —que pueden revelar desde predisposiciones genéticas hasta comportamientos íntimos— exige mecanismos robustos de privacidad diferencial, creando una tensión aparentemente irresoluble entre explicabilidad y confidencialidad.

El sector financiero ofrece otro escenario paradigmático donde esta tensión adquiere dimensiones socioeconómicas significativas. Los sistemas algorítmicos de evaluación crediticia, como el FICO Score en Estados Unidos o el Buró de Crédito en México, han evolucionado desde

[12] GULSHAN, V; PENG, L; CORAM, M; STUMPE, M. C.; WU, D & NARAYANASWAMY, A. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Journal of the American Medical Association, 316 (22), pp. 2402-2410.

simples modelos estadísticos hasta complejos sistemas de aprendizaje automático que analizan cientos de variables para determinar la solvencia de un solicitante. Como señalan Martínez-Plumed et al. (2024), estos algoritmos frecuentemente "perpetúan y amplifican desigualdades históricas al incorporar variables proxy que correlacionan con características protegidas como raza o género"[13].

La forma en que los sistemas de evaluación crediticia funcionan y han evolucionado a lo largo del tiempo es compleja. Desde que se fundó la primera agencia de informes crediticios en EE. UU. Se comenzó utilizando métodos en donde se avaluaban los informes, se clasificaba el riesgo crediticio y se modelaba la información estadísticamente en modelos lineales en el que usaban información como socioeconómica general, demográfica y de ingresos variables, utilizando las suposiciones en el que el comportamiento pasado del consumidor es un indicador del comportamiento futuro[14]. Mientras que, en los actuales métodos de evaluación crediticia utilizando aprendizaje automatizado se utilizan modelos basados en árboles en donde se utilizan cientos de variables en tiempo real para predecir la capacidad de pago de una persona, posibilitando, por un lado, una predicción más precisa, pero al mismo tiempo deja un posible sesgo algorítmico donde se trata de evocar la explicabilidad y transparencia de los datos obtenidos.

Esta situación evoca la parábola kafkiana del guardián de la ley: ciudadanos que permanecen eternamente a las puertas del sistema financiero, sin comprender las reglas invisibles que determinan su exclusión. Paradójicamente, las técnicas de privacidad diferencial, al "difuminar" los datos individuales, pueden simultáneamente proteger la privacidad de los consumidores mientras oscurecen aún más los

[13] ADEYELU, O. O; UGOCHUKWU, C. E. & SHONIBARE, M. A. (2024). Ethical implications of AI in financial decision-making: A review with real-world applications. *International Journal of Applied Research in Social Sciences, 6* (4), pp. 608–630.

[14] MALEMPATI, M. (2025). The evolution of credit data and the role of machine learning in modern credit scoring. In *The Intelligent Ledger: Harnessing Artificial Intelligence, Big Data, and Cloud Power to Revolutionize Finance, Credit, and Security* (pp. 19-35). Deep Science Publishing. https://doi.org/10.70593/978-93-49910-16-4_2

patrones discriminatorios subyacentes, creando una "opacidad de segundo orden" que dificulta la auditoría de sesgos algorítmicos [15].

Quizás el caso más emblemático de esta tensión se manifiesta en el sector judicial, donde algoritmos como COMPAS están siendo implementados para tareas tan sensibles como la evaluación de riesgos procesales o la recomendación de sentencias. La investigación de ProPublica reveló que este sistema "clasificaba incorrectamente a acusados afroamericanos como "alto riesgo" casi el doble de veces que, a acusados caucásicos, evidenciando sesgos sistémicos"[16]. Sin embargo, la empresa desarrolladora, Northpointe (ahora Equivant), se negó a revelar completamente el funcionamiento interno del algoritmo, argumentando protección de propiedad intelectual. Este caso ilustra vívidamente la colisión entre el derecho empresarial a proteger secretos comerciales y principios fundamentales del debido proceso como el derecho a confrontar la evidencia adversa. Es comparable a un sistema judicial medieval donde las pruebas determinantes permanecían selladas en cofres inaccesibles para el acusado, con la particularidad de que, en la era digital, estos "cofres algorítmicos" están protegidos no solo por cerraduras físicas sino por capas de complejidad matemática y protecciones legales de propiedad intelectual.

IV. Soluciones Intermedias: Hacia una Conciliación Pragmática

Ante esta aparente contradicción entre valores igualmente importantes, diversas aproximaciones técnicas, legales y procedimentales ofrecen vías intermedias prometedoras. En primer lugar se encuentra la "Explicabilidad por capas". La aparente contradicción entre la necesaria transparencia de sistemas algorítmicos que toman decisiones críticas y la imprescindible protección de datos personales no constituye necesariamente un dilema irresoluble. Al igual que en la arquitectura de seguridad nuclear se desarrollan "sistemas de defensa en profundidad" con múltiples capas de protección, el concepto de "explicabilidad

[15] JOSHUA, A. Kroll; HUEY, Joanna; BAROCAS, Solon; FELTEN, Edward W; REIDENBERG, Joel R; ROBINSON, David G & HARLAN, Yu Accountable Algorithms, 165 U. Pa. L. Rev. 633 (2017).

[16] ANGWIN, J; LARSON, J; MATTU, S & KIRCHNER, L. (2024). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica.

por capas" propone un enfoque estratificado que modula el nivel de transparencia según el contexto, necesidad y audiencia legítima [17].

Este modelo puede estructurarse como un sistema concéntrico de acceso diferenciado a la información algorítmica. En el núcleo, encontraríamos una "capa técnica profunda" reservada para auditorías técnicas independientes y reguladores especializados, bajo condiciones estrictas de confidencialidad. Esta capa permitiría el escrutinio completo del código, parámetros y conjuntos de entrenamiento del algoritmo, similar al acceso que tienen los inspectores del Organismo Internacional de Energía Atómica (OIEA) a instalaciones nucleares sensibles: completo, pero estrictamente controlado mediante protocolos de seguridad multinivel.

Rodeando este núcleo, una "capa procesal intermedia" estaría dirigida específicamente a los sujetos directamente afectados por decisiones algorítmicas. Por ejemplo, un solicitante de crédito rechazado recibiría una explicación personalizada que identifica los factores específicos que determinaron su caso particular (como relación deuda-ingresos, historial de pagos atrasados o duración del historial crediticio), sin necesidad de comprender la compleja matemática del modelo subyacente. Esta aproximación se alinea con lo que Wachter *et al*. (2023)[18] denominan "Explicaciones Contrafácticas": respuestas a la pregunta "¿qué tendría que cambiar en mi situación para obtener un resultado diferente?".

Finalmente, la capa más externa, o "capa conceptual superficial", estaría orientada al público general, ofreciendo una comprensión accesible de los principios que rigen el sistema, sus objetivos declarados, limitaciones conocidas y mecanismos de supervisión. Esta capa funcionaría como un "prospecto algorítmico" comparable a los prospectos farmacéuticos que informan sobre medicamentos sin revelar la fórmula química exacta, permitiendo un escrutinio público significativo sin comprometer secretos comerciales ni facilitar la manipulación malintencionada del sistema.

---

[17] CHEN, L; CRUZ, I; SHERIFF, S & KAMAR, E. (2024). Layered explanations: A practical framework for AI transparency. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 134-145.

[18] WACHTER, S; MITTELSTADT, B & RUSSELL, C. (2023). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology, 36 (2), pp. 841-887.

Esta aproximación estratificada puede implementarse técnicamente mediante diversos métodos de "interpretabilidad *post-hoc*" como los mapas de saliencia en redes neuronales convolucionales o los modelos sustitutos interpretables (*surrogate models*), que generan aproximaciones simplificadas pero fieles de sistemas más complejos. La ventaja fundamental de este enfoque es que permite equilibrar la rendición de cuentas democrática con la protección de intereses legítimos de privacidad y propiedad intelectual, adaptando el nivel de explicabilidad al contexto específico y a la necesidad concreta de cada parte interesada.

Por otra parte, el desafío de proteger datos sensibles sin sacrificar la utilidad analítica o la transparencia algorítmica ha estimulado el desarrollo de técnicas avanzadas que podríamos denominar "privacidad por diseño". Estas aproximaciones van más allá de la mera anonimización superficial, incorporando protecciones de privacidad en la arquitectura misma de los sistemas algorítmicos, comparable a cómo los principios de seguridad pasiva se incorporan estructuralmente en el diseño arquitectónico de edificios resistentes a sismos.

El "aprendizaje federado" representa una revolución paradigmática en este ámbito, comparable al paso de sistemas centralizados de agua potable a sistemas de captación pluvial distribuidos. En lugar de concentrar datos sensibles en servidores centrales vulnerables a brechas de seguridad, esta metodología desarrollada inicialmente por Google[19] permite entrenar algoritmos de forma distribuida manteniendo los datos personales en los dispositivos de los usuarios y compartiendo únicamente modelos agregados. Un ejemplo concreto lo encontramos en la aplicación *SwiftKey*, que mejora sus predicciones de texto analizando patrones de escritura individuales sin jamás transmitir el contenido exacto de los mensajes a servidores centrales[20].

---

[19] LI, T; SAHU; A. K., TALWALKAR, A & SMITH, V. (2024). Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Processing Magazine, 41 (3), pp. 148-170

[20] Microsoft. (s. f.). Microsoft SwiftKey Keyboard: Privacy Questions and your Data. Microsoft Support. Recuperado el 9 de mayo de 2025, de https://support.microsoft.com/en-us/topic/microsoft-swiftkey-keyboard-privacy-questions-and-your-data-07e13677-6b38-4ad0-bad0-d41207cab6de

Complementariamente, la técnica de "distilación de conocimiento", conceptualizada originalmente por Geoffrey Hinton[21], posibilita crear modelos simplificados que capturan la funcionalidad esencial de sistemas más complejos sin incorporar datos sensibles del modelo original. Esta aproximación es comparable a cómo un maestro transmite conocimiento esencial a un alumno sin revelar todas sus fuentes originales o experiencias personales.

Asimismo, la "computación multipartita segura" ofrece un enfoque prometedor para escenarios donde múltiples entidades necesitan computar conjuntamente funciones sobre sus datos privados sin revelarlos entre sí. Esta metodología, fundamentada en resultados teóricos de criptografía moderna, es comparable a una mesa de póker virtual donde cada jugador puede verificar que el juego es justo sin ver las cartas de los demás.

Estas técnicas, al incorporar la privacidad en la arquitectura misma de los sistemas, permiten trascender la aparente dicotomía entre transparencia y confidencialidad, creando lo que podríamos denominar "sistemas de transparencia selectiva" que revelan su funcionamiento general sin comprometer datos sensibles específicos.

Sin embargo, existen diversos marcos regulatorios equilibrados que ofrecen una perspectiva jurídica.

La conciliación entre transparencia algorítmica y privacidad diferencial requiere marcos regulatorios que superen el enfoque binario (todo público versus todo confidencial) para adoptar aproximaciones matizadas y contextuales. Tres principios fundamentales podrían orientar esta regulación equilibrada.

Primero, la proporcionalidad contextualizada: el nivel de transparencia requerido debe ser proporcional a la sensibilidad de la decisión algorítmica y su impacto en derechos fundamentales. Una aproximación de este tipo la encontramos en el recién aprobado *AI Act* europeo (2024), que establece "categorías diferenciadas de sistemas algorítmicos según su nivel de riesgo, con requisitos de transparencia progresivamente más

[21] HINTON, G; VINYALS, O & DEAN, J. (2023). Distilling the Knowledge in a Neural Network. NIPS Deep Learning Workshop, 23 (4), pp. 1-9.

estrictos conforme aumenta el impacto potencial"[22]. Así, mientras que un algoritmo de recomendación musical estaría sujeto a requisitos mínimos de explicabilidad, un sistema de evaluación crediticia o diagnóstico médico enfrentaría estándares más rigurosos. Esta aproximación escalonada evita tanto la sobrerregulación que podría asfixiar la innovación como la infra regulación que dejaría desprotegidos derechos fundamentales.

Segundo, la rendición de cuentas por diseño: los sistemas algorítmicos deberían ser conceptualizados desde su fase inicial para facilitar la auditoría externa sin comprometer datos sensibles. Este principio, inspirado en el concepto de "*privacy by design*" popularizado por Ann Cavoukian, requeriría que los desarrolladores implementen mecanismos de "trazabilidad selectiva" que permitan reconstruir el proceso decisorio en casos específicos sin exponer el conjunto completo de datos o parámetros. La implementación más avanzada de este concepto la encontramos en el "*Algorithmic Impact Assessment*" obligatorio para todas las decisiones automatizadas en el sector público canadiense desde 2021, que requiere "documentar exhaustivamente las consideraciones de diseño, fuentes de datos y mecanismos de mitigación de sesgos antes de la implementación de cualquier sistema algorítmico de toma de decisiones" (Government of Canada, 2023)[23].

Tercero, la supervisión independiente especializada: la creación de órganos técnicos con competencias específicas para auditar algoritmos críticos bajo estrictas normas de confidencialidad.

La experiencia internacional más cercana a este modelo la encontramos en el *Norsk Råd For Digital Etikk* (NORDE) con su proyecto de investigación "*Etisk Risikovurdering av KI i Praksis*" (ENACT), lanzado en 2023. ENACT

[22] European Parliament and Council. (2024, 12 de julio). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689.

[23] Government of Canada. (s. f.). Algorithmic Impact Assessment tool. Treasury Board of Canada Secretariat. Recuperado el 9 de mayo de 2025, de https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

desarrolla una metodología de "trazabilidad selectiva" que permite reconstruir decisiones de sistemas de IA de alto riesgo sin exponer la totalidad de los datos o parámetros internos, mediante acuerdos de confidencialidad con expertos externos y garantizando así tanto la protección de la propiedad intelectual como la posibilidad de auditorías independientes [24].

Un enfoque regulatorio que integre estos tres principios permitiría equilibrar la necesaria rendición de cuentas democrática con la protección de intereses legítimos como la privacidad individual, la seguridad de los sistemas y los derechos de propiedad intelectual, reconociendo que la transparencia no es un valor absoluto sino contextual, cuya implementación debe adaptarse a las circunstancias específicas de cada caso.

V. Implicaciones para el Estado de Derecho en la Era Digital

La tensión entre transparencia algorítmica y privacidad diferencial trasciende lo meramente técnico para situarse en el núcleo de cómo conceptualizamos el Estado de Derecho en sociedades cada vez más mediadas por algoritmos. Podemos identificar al menos tres dimensiones fundamentales. En primer lugar, se encuentra la reconfiguración del principio de publicidad de las normas y actuaciones estatales, el cual es pilar fundamental del Estado de Derecho moderno, y enfrenta una transformación radical. Si, como afirmaba Bobbio[25], "la democracia es el gobierno del poder público en público" ¿cómo reconciliar este principio con la creciente delegación de decisiones públicas a sistemas algorítmicos opacos?

La respuesta podría encontrarse en lo que podríamos denominar una "publicidad adaptativa": un modelo que garantice la visibilidad y comprensibilidad de las decisiones algorítmicas para los ciudadanos afectados, aunque el mecanismo técnico específico permanezca parcialmente reservado por razones de seguridad, eficacia o protección de datos.

---

[24] Norsk råd for digital etikk. (2023). Etisk risikovurdering av KI i praksis. NORDE. Recuperado el 9 de mayo de 2025, de https://www.norde.digital/

[25] BOBBIO, N. (2013). *Democracia y secreto*. Fondo de Cultura Económica, pp. 37-38.

En segundo lugar, se encuentra el derecho al debido proceso, entendido tradicionalmente como la posibilidad de conocer y controvertir las pruebas y razonamientos que fundamentan una decisión, requiere una reconceptualización en entornos algorítmicos. El desafío consiste en definir qué constituye una "explicación suficiente" cuando el proceso decisorio involucra sistemas de aprendizaje automático con millones de parámetros.

Un enfoque prometedor podría ser el desarrollo de "*proxies* explicativos": representaciones simplificadas pero fieles del razonamiento algorítmico, que permitan a ciudadanos sin conocimientos técnicos avanzados comprender y potencialmente impugnar decisiones automatizadas que les afectan.

Finalmente, la equidad procesal —entendida como igualdad de armas entre las partes— adquiere nuevas dimensiones cuando una de las "partes" es un algoritmo. La asimetría de conocimiento entre los desarrolladores del sistema y los ciudadanos afectados por sus decisiones crea un desequilibrio fundamental que amenaza la legitimidad del proceso.

Para mitigar esta asimetría, podría contemplarse el desarrollo de "defensores algorítmicos públicos": expertos técnicos independientes que asistan a ciudadanos en la comprensión e impugnación de decisiones automatizadas, similar a cómo los defensores públicos proporcionan asistencia legal a quienes no pueden costearla.

VI. Conclusiones: Hacia un Nuevo Equilibrio

La opacidad algorítmica, como hemos demostrado, compromete principios fundacionales del Estado de Derecho. Cuando COMPAS clasifica a un acusado como "alto riesgo" de reincidencia sin revelar su metodología, o cuando un algoritmo crediticio rechaza un préstamo sin explicación substantiva, estamos ante un retroceso inquietante a formas pre-modernas de poder: decisiones que afectan derechos fundamentales pero que permanecen inmunes al escrutinio racional.

Por otra parte, la privacidad diferencial ofrece una respuesta técnicamente sofisticada al problema de protección de datos, pero genera sus propias problemáticas.

Las aproximaciones intermedias que hemos identificado —explicabilidad por capas, técnicas avanzadas de privacidad, y marcos regula-

torios equilibrados— ofrecen caminos prometedores, pero requieren una implementación cuidadosamente calibrada. La propuesta de "explicabilidad por capas" encuentra un paralelismo interesante en el principio de "defensa en profundidad" utilizado en seguridad nuclear: múltiples barreras de protección con diferentes funciones y niveles de acceso.

En el ámbito regulatorio, el recién aprobado *AI Act* europeo (2024) representa el intento más ambicioso hasta la fecha de abordar esta tensión. Su enfoque basado en riesgos, que establece requisitos de transparencia proporcionalmente más estrictos según el impacto potencial del sistema algorítmico, ofrece un marco conceptual valioso. Sin embargo, este modelo enfrenta desafíos significativos en su implementación práctica: la categorización de sistemas por nivel de riesgo presupone una capacidad evaluativa que, paradójicamente, requiere cierto grado de transparencia previa. Estamos ante lo que podríamos denominar una "circularidad regulatoria": para determinar qué nivel de transparencia exigir a un algoritmo, necesitamos primero comprender su funcionamiento con suficiente detalle para evaluar sus riesgos potenciales.

Las implicaciones para el Estado de Derecho en la era digital son profundas y multifacéticas. La reconfiguración del principio de publicidad requiere lo que podríamos denominar una "transparencia adaptativa": no necesariamente la exposición completa de cada parámetro técnico, sino la garantía de que las decisiones algorítmicas sean comprensibles y contestables por los ciudadanos afectados.

La búsqueda de equilibrio entre transparencia algorítmica y privacidad diferencial es, en última instancia, una manifestación específica de la tensión perenne entre valores democráticos fundamentales. Al igual que la libertad de expresión debe equilibrarse con la protección de otros derechos, o la seguridad nacional con las libertades civiles, la sociedad digital deberá encontrar su propio punto de equilibrio entre la necesaria transparencia de los sistemas que toman decisiones críticas y la imprescindible protección de la privacidad individual en una era de dataficación masiva y capitalismo de vigilancia.

El desafío para juristas, tecnólogos, filósofos y ciudadanos no consiste meramente en encontrar un punto de equilibrio momentáneo, sino en desarrollar marcos conceptuales, técnicos y normativos que permitan una recalibración continua y responsiva. En palabras de Lawrence Lessig, "el código es ley" —la arquitectura tecnológica incorpora valores

y distribuye poder— pero, complementariamente, necesitamos asegurar que la ley pueda efectivamente gobernar el código, estableciendo los principios fundamentales que deben orientar el diseño tecnológico en una sociedad democrática.

Al final, la tensión aparentemente irresoluble entre algoritmos opacos y privacidad diferencial nos recuerda que la tecnología nunca es neutral. Cada decisión de diseño algorítmico, cada implementación de mecanismos de privacidad, cada marco regulatorio refleja y refuerza determinadas concepciones sobre la relación entre individuo, sociedad y Estado. En la era de la hiperconectividad digital, estos sistemas técnicos se convierten en infraestructuras de gobernanza de facto, con profundas implicaciones para el ejercicio de derechos fundamentales y la distribución de poder social.

Es alentador observar cómo, mientras debatimos estos desafíos aparentemente irresolubles a nivel teórico, la industria tecnológica ya avanza en implementaciones concretas que buscan el equilibrio que hemos analizado. En octubre de 2024, Google liberó PipelineDP4J, una biblioteca de código abierto que implementa privacidad diferencial en Java, extendiendo sus soluciones de privacidad mejorada a "más de la mitad de los desarrolladores a nivel mundial" al incorporar ahora Python, Java, Go y C++ entre sus opciones[26]. Lo particularmente significativo de esta iniciativa no es solo la democratización de estas herramientas matemáticas sofisticadas, sino su enfoque dual: por un lado, facilita el acceso a técnicas que protegen datos individuales mientras permiten análisis agregados valiosos; por otro, proporciona mediante DP-Auditorium mecanismos para verificar independientemente las garantías de privacidad ofrecidas, sin necesidad de acceso a las propiedades internas de las aplicaciones. Este desarrollo cristaliza perfectamente el equilibrio dinámico entre utilidad de los datos y protección de la privacidad que hemos propuesto como fundamento para reconciliar valores aparentemente antagónicos. Más que un mero avance técnico, representa la materialización práctica de los principios de "transparencia selectiva" y "rendición de cuentas por diseño" que hemos identificado como esen-

[26] GUEVARA, M. (2024, 31 de octubre). Open sourcing our Java-based differential privacy library. Google Privacy & Security Blog. Recuperado el 12 de mayo de 2025, de https://blog.google/technology/safety-security/open-sourcing-our-java-based-differential-privacy-library/

ciales. Demuestra que el camino hacia la coexistencia de algoritmos auditables y privacidad robusta no es una aspiración abstracta, sino una realidad emergente impulsada por implementaciones concretas que permiten extraer valor analítico de los datos sin comprometer derechos fundamentales de privacidad.

La disyuntiva no es, por tanto, meramente técnica sino profundamente política y ética: ¿qué tipo de sociedad queremos construir a través de nuestras decisiones tecnológicas? Una respuesta responsable exige trascender tanto el determinismo tecnológico como la regulación reactiva, adoptando en su lugar un enfoque de "gobernanza anticipatoria" que integre consideraciones éticas, sociales y democráticas desde las etapas iniciales del diseño tecnológico. Solo así podremos asegurar que la revolución algorítmica fortalezca, en lugar de socavar, los principios fundamentales del Estado de Derecho y la dignidad humana en el siglo XXI.