Daimon. Revista Internacional de Filosofía, en prensa, aceptado para publicación tras revisión por pares doble ciego.
ISSN: 1989-4651 (electrónico) http://dx.doi.org/10.6018/daimon.682291

Autoridad epistémica algorítmica ante la razón pública

Algorithmic Epistemic Authority and Public Reason

FERNANDO A. RAMOS-ZAGA [1]

Resumen: La incorporación de sistemas algorítmicos en la toma de decisiones públicas ha modificado las condiciones de ejercicio de la autoridad administrativa. Cuando sus resultados operan como razones excluyentes para la acción, la opacidad técnica y justificatoria limita la evaluación y plantea problemas de legitimidad democrática. El artículo analiza las condiciones normativas de legitimidad de la autoridad epistémica atribuida a sistemas algorítmicos en la toma de decisiones públicas. Se sostiene que la fiabilidad predictiva y la explicabilidad técnica no bastan para legitimar dicha deferencia. Sobre esta base, se formula la desconfianza lúcida como una disposición epistemológico-política que reconoce el derecho ciudadano a suspender la aceptación de decisiones carentes de justificación pública. Se concluye que la autoridad algorítmica debe subordinarse a razones comprensibles, discutibles y revisables, y que la administración debe asumir la carga de su justificación.

Palabras clave: deferencia práctica, democracia deliberativa, escrutinio público, legitimidad democrática, transparencia algorítmica.

Abstract: The adoption of algorithmic systems in public decision-making has reshaped the conditions under which administrative authority is exercised. When their outputs function as exclusionary reasons for action, technical and justificatory opacity constrains evaluation and raises challenges to democratic legitimacy. This article examines the normative conditions governing the legitimacy of epistemic authority attributed to algorithmic systems in public decision-making. It argues that predictive reliability and technical explainability are insufficient to justify such deference. On this basis, lucid distrust is advanced as an epistemic-political disposition that recognizes citizens' right to withhold acceptance of decisions lacking public justification. It concludes that algorithmic authority must remain subordinate to intelligible, contestable, and revisable reasons, and that public authorities must bear the burden of providing such justification.

Keywords: algorithmic transparency, deliberative democracy, democratic legitimacy, practical deference, public scrutiny.

1. Introducción

La integración progresiva de sistemas de aprendizaje automático en las instituciones estatales ha transformado la mecánica tradicional del poder público en las democracias contemporáneas. Los algoritmos de predicción y clasificación no operan como meras

Recibido: 02/10/2025. Aceptado: 07/08/2026.
[1] Docente investigador. Sus principales líneas de investigación abarcan la bioética, la filosofía del derecho, la regulación de nuevas tecnologías y economía del comportamiento. Universidad Privada del Norte, Perú. ORCID: https://orcid.org/0000-0001-6301-9460 Mail de contacto: fernandozaga@gmail.com

herramientas auxiliares de cálculo. También estructuran las premisas decisorias, fijan criterios normativos y condicionan el ejercicio de la potestad administrativa, con efectos directos sobre los derechos ciudadanos (Hildebrandt, 2015, p. 185). La automatización del juicio institucional desplaza la apreciación discrecional humana hacia artefactos informáticos y convierte lo que parecía un ajuste de eficiencia técnica en un problema de legitimidad política.

En el ámbito de la teoría normativa, la adscripción de funciones decisorias a estos sistemas remite al concepto de autoridad práctica, entendida como la capacidad de emitir directivas que funcionan como razones excluyentes para la acción (Raz, 1986, p. 41). En el plano del conocimiento, la autoridad epistémica exige que el sujeto sustituya sus propias razones por el dictamen de la fuente, siempre que una evaluación reflexiva justifique esa cesión del juicio (Zagzebski, 2012, p. 107). Sin embargo, la opacidad constitutiva de los modelos de aprendizaje profundo, en sus dimensiones epistémica, semántica y justificatoria, altera el presupuesto reflexivo de la deferencia (Hauswald, 2025, p. 719) e instaura un acatamiento a la fuente que carece de la validación crítica exigida por la teoría de la autoridad.

A pesar del despliegue de los marcos de gobernanza centrados en la transparencia formal, la inteligencia artificial explicable y la equidad algorítmica, persiste un vacío conceptual sobre los criterios normativos que hacen justificable la deferencia práctica en el ámbito público. Las soluciones técnicas disponibles resultan insuficientes porque confunden la explicabilidad estadística con la fundamentación normativa (Zednik y Verreault-Julien, 2026, p. 123), mientras que las métricas formales de equidad ocultan elecciones axiológicas no deliberadas (Campione et al., 2026, p. 3). Además, los mecanismos de contestabilidad *ex post* invierten de manera indebida la carga de la prueba sobre el ciudadano (Henin y Le Métayer, 2021, p. 8; Simbeck, 2022, p. 95). En consecuencia, falta un estándar ético-político que permita determinar cuándo una decisión algorítmica puede ser aceptada de manera justificada por la comunidad cívica.

Las consecuencias de esta laguna teórica y normativa desbordan el plano procedimental e inciden directamente en la estructura básica de la sociedad. La delegación continuada en modelos estocásticos opacos introduce distorsiones acumulativas que degradan la justicia de trasfondo (Rawls, 1993, p. 257) e imponen formas de dominación burocrática no sujetas a control democrático (Frost, 2024, p. 783). Tal dinámica no solo erosiona el deber de la administración de ofrecer razones públicas, sino que también favorece la descualificación

profesional de los funcionarios y la progresiva alienación del ciudadano frente a las decisiones institucionales.

En este marco, el objetivo del artículo es analizar las condiciones normativas que legitiman la autoridad epistémica atribuida a sistemas algorítmicos opacos en la toma de decisiones públicas. La relevancia de este propósito reside en ofrecer una vía de análisis que supera tanto el optimismo tecnológico como el rechazo tecnófobo. Con ello, se propone el concepto de desconfianza lúcida como una disposición epistemológico-política normativamente fundamentada para salvaguardar la autonomía cívica y la legitimidad democrática.

2. Pretensión de autoridad y deferencia algorítmica

Los sistemas algorítmicos empleados en la decisión pública no se limitan a procesar información o a generar recomendaciones consultivas, sino que progresivamente son investidos institucionalmente de funciones que presentan los rasgos formales de la autoridad.

Al respecto, la autoridad práctica, en su acepción más rigurosa, designa la capacidad normativa de alterar la situación obligacional de los agentes mediante directivas que constituyen razones protegidas para actuar (Raz, 1979, p. 18). Una razón protegida articula una razón de primer orden para ejecutar un acto con una razón excluyente de segundo orden que desestima los motivos para no ejecutarlo. De modo que la aceptación de una autoridad implica el compromiso de seguir sus mandatos con independencia de la propia evaluación sobre los méritos del caso, lo cual supone una cesión del juicio que no se fundamenta en el contenido de la determinación, sino en el carácter de la fuente que la emite (Raz, 1986, p. 39).

Sin embargo, esta pretensión de autoridad no exige la posesión de una mente consciente por parte de la entidad que la reclama. La agencia artificial permite concebir entidades construidas capaces de percibir su entorno y actuar en él de forma autónoma sin intervención directa de su creador (Hildebrandt, 2015, p. 21). Ahora bien, la cuestión relevante no es si los sistemas algorítmicos poseen agencia en sentido fuerte, sino si sus salidas operan funcionalmente como razones excluyentes en el razonamiento práctico de quienes los emplean. En la medida en que un operador institucional asume la estimación algorítmica como premisa decisoria sin evaluar de manera independiente las razones subyacentes al caso, se reproduce la mecánica de la razón excluyente con independencia de que la fuente sea humana o artificial.

En el ámbito de la epistemología social, esta equiparación funcional adquiere una precisión decisiva. La autoridad epistémica designa el poder normativo cuya directiva se adopta como

razón para creer o actuar, de tal modo que el hecho de que la autoridad sostenga una creencia reemplaza, en lugar de sumarse a, las razones del sujeto en el dominio correspondiente (Zagzebski, 2012, p. 107). Sin embargo, la justificación de esta razón excluyente descansa en el juicio escrupuloso del sujeto acerca de que es más probable alcanzar la verdad o evitar el error al deferir que al indagar por cuenta propia (Zagzebski, 2012, p. 110). Por tanto, la deferencia epistémica legítima presupone una evaluación reflexiva sobre la calidad de la fuente, presupuesto que la opacidad algorítmica compromete de raíz.

Esta opacidad estructural presenta una complejidad que trasciende la mera dificultad técnica. Cabe distinguir, al menos, tres niveles de inaccesibilidad. La opacidad epistémica impide al sujeto conocer el valor de verdad de la proposición generada por el sistema. La opacidad semántica dificulta la comprensión de las condiciones bajo las cuales la proposición sería verdadera. La opacidad justificatoria hace imposible reconstruir las razones que sustentan el resultado (Hauswald, 2025, p. 719). Esta estratificación revela que no se trata simplemente de un déficit de información subsanable mediante mejores interfaces de usuario, sino de una condición estructural derivada de la naturaleza misma de los modelos estocásticos de alta dimensionalidad.

Lejos de constituir un subproducto contingente, dicha opacidad resulta un elemento inherente al paradigma computacional dominante. Los modelos de aprendizaje profundo operan mediante la aproximación de gradientes con ruido, mutaciones estocásticas y optimizaciones estadísticas multietapa de hiperparámetros (Wang et al., 2023, p. 2). En consecuencia, sus salidas no se derivan de deducciones normativas o de explicaciones causales inteligibles, sino de correlaciones estadísticas extraídas de conjuntos de datos cuya representatividad y calidad resultan, a su vez, difícilmente verificables. Tal condición impide que el operador humano pueda efectuar el juicio reflexivo escrupuloso que la teoría de la autoridad epistémica exige como presupuesto de la deferencia legítima.

Más allá de la opacidad técnica, la pretensión de autoridad algorítmica se apoya en una apariencia de neutralidad valorativa que resulta seriamente cuestionable. Los datos de entrenamiento no son registros neutrales de la realidad, sino el resultado de procesos de discretización y traducción que sesgan de forma inherente la representación del fenómeno (Hildebrandt, 2015, p. 36). Los datos masivos extraídos de Internet sobrerepresentan a usuarios jóvenes, masculinos y de países desarrollados, reificando visiones hegemónicas y amplificando sesgos discriminatorios (Bender et al., 2021, p. 613). Así pues, la ilusión de objetividad técnica

encubre opciones de diseño, elecciones de parámetros y selecciones de datos que reflejan concepciones valorativas y prioridades políticas no explicitadas.

Dentro de la arquitectura institucional, la incorporación de estas estimaciones algorítmicas consolida modalidades específicas de deferencia. Cuando un sistema de evaluación de riesgo penal clasifica a un acusado como de alta peligrosidad, esa clasificación opera funcionalmente como una razón predeterminada para el juez, cuyo margen de apreciación discrecional queda condicionado por la estimación del sistema (Ozer et al., 2024, p. 201). Con todo, la evidencia experimental muestra que los ciudadanos no eximen de responsabilidad al funcionario por alinearse con el algoritmo, sino que interpretan tal alineamiento como una abdicación del deber de ejercer el propio juicio (Ozer et al., 2024, p. 209). De este modo, la deferencia algorítmica genera una tensión entre la presión institucional a conformarse con la recomendación automatizada y la expectativa cívica de que la autoridad humana preserve su capacidad deliberativa.

Más significativamente, la dimensión anticipatoria de la gobernanza algorítmica agrava esta tensión deliberativa. A diferencia del derecho, que apela a la persona como sujeto capaz de asumir responsabilidad por sus actos, la gobernanza algorítmica anticipa e inhabilita preventivamente la voluntad humana mediante dinámicas invisibles de clasificación y predicción (Hildebrandt, 2015, p. 185). Este desplazamiento transforma la función del derecho, puesto que la razón excluyente de la intención del ciudadano sustituye la orientación normativa por la gestión estocástica del riesgo poblacional. Bajo esta lógica, el sujeto deja de ser interpelado como agente moral y es reducido a un perfil estadístico cuya conducta futura se estima mediante correlaciones pasadas.

Así articulada, la deferencia algorítmica adquiere relevancia normativa cuando las salidas del sistema operan como razones excluyentes sin que el decisor pueda evaluar de manera independiente su fundamento. La opacidad estructural, la incorporación de sesgos valorativos y las presiones institucionales impiden equiparar la capacidad predictiva con una autoridad legítima. Por ello, la sección siguiente examina el principio de razón pública como marco para determinar en qué condiciones la cesión del juicio ante sistemas algorítmicos puede considerarse justificable.

3. Razón pública y justificabilidad de la deferencia algorítmica

La cuestión de la justificabilidad de la deferencia práctica hacia sistemas algorítmicos no puede resolverse apelando exclusivamente a criterios epistemológicos de fiabilidad o precisión

predictiva, sino que exige un principio normativo capaz de articular las dimensiones epistémica y política de la legitimidad.

El principio de razón pública ofrece precisamente este marco normativo, dado que establece las condiciones bajo las cuales las decisiones coercitivas del Estado resultan aceptables para todos los ciudadanos concebidos como libres e iguales (Rawls, 1993, p. 213). La razón pública es la razón propia de los ciudadanos en un régimen democrático cuando deliberan sobre esenciales constitucionales y cuestiones de justicia básica, y sus contenidos deben fundarse en valores políticos compartidos con independencia de doctrinas comprensivas particulares (Rawls, 1993, p. 212).

En el núcleo operativo de este principio se ubica el criterio de reciprocidad. Toda propuesta de principio o política pública debe poder explicarse mediante razones normativas e informacionales que cada ciudadano pueda aceptar razonablemente (Rawls, 1993, p. 226). Por ello, cuando un sistema algorítmico ejerce funciones de autoridad en el ámbito de la administración pública, mediante decisiones que inciden directamente sobre derechos fundamentales, la distribución de bienes primarios o el ejercicio de la coerción estatal, la pretensión de legitimidad de sus determinaciones queda sometida a esta exigencia.

Bajo este criterio, una estimación estadística inaccesible para la deliberación intersubjetiva no puede funcionar como razón pública válida para el ejercicio de la coerción o la distribución de recursos. En consecuencia, el deber de civilidad, entendido como la obligación moral de explicarse recíprocamente por las decisiones fundamentales apelando a los valores de la razón pública (Rawls, 1993, p. 219), se extiende a los actos institucionales sustentados en determinaciones algorítmicas.

Sin embargo, esta exigencia de razón pública no se limita a la mera disponibilidad formal de información sobre el funcionamiento del sistema. La inteligibilidad exigida es de carácter normativo, no técnico. Ello significa que no basta con que el código fuente sea accesible o con que se ofrezca una descripción simplificada del algoritmo, sino que las premisas decisorias del sistema deben ser traducibles a razones comprensibles y evaluables dentro de los cánones de la argumentación pública. La distinción entre razones motivacionales, explicativas y normativas resulta aquí decisiva. Las razones motivacionales son los patrones o atributos ponderados por el sistema para generar su salida. Las razones explicativas son los estados y procesos computacionales que causan la decisión. Las razones normativas son los estándares éticos, legales o epistémicos externos que determinan si la decisión es correcta (Zednik y Verreault-

Julien, 2026, p. 110). La razón pública opera precisamente en el dominio de las razones normativas, y la explicabilidad técnica, por sofisticada que sea, no puede suplir su ausencia.

A la luz de esta distinción normativa, la adecuación epistémica de los presupuestos del sistema algorítmico constituye una condición necesaria, aunque no suficiente, de la justificabilidad de la deferencia. La concepción de servicio de la autoridad exige que toda directiva autoritativa se fundamente en las razones subordinadas que ya se aplican de manera independiente a los gobernados (Raz, 1986, p. 46). Dicho de otro modo, la autoridad no crea deberes completamente nuevos, sino que facilita el cumplimiento de las razones que ya obligan al sujeto. Si los datos de entrenamiento del sistema incorporan correlaciones espurias, si la función de optimización pondera variables irrelevantes o si la arquitectura del modelo genera sesgos sistemáticos, entonces la directiva algorítmica no refleja las razones subyacentes al caso, con lo cual su pretensión de autoridad carece de fundamento.

Sin embargo, la verificación de dicha adecuación epistémica no puede efectuarse exclusivamente mediante procedimientos internos al propio sistema. La circularidad epistémica que supone utilizar las mismas funciones que calcularon la salida para justificar la creencia en ella confirma la robustez técnica del procedimiento, pero no su validez normativa (Durán, 2026, p. 62). Por tanto, se requiere un *test* de evaluación pública fundado en la escrutabilidad de los presupuestos del sistema y en la revisabilidad deliberativa de sus ponderaciones axiológicas.

Dicho *test* de evaluación exige, como mínimo, que sea posible identificar las concepciones valorativas incorporadas en el diseño del sistema, que los datos de entrenamiento sean documentables y auditables, y que las ponderaciones entre valores en conflicto puedan ser sometidas a deliberación democrática. Tal exigencia no es trivial, ya que la recolección indiscriminada de datos masivos genera una "deuda de documentación" que hace imposible auditar retroactivamente los sesgos codificados en la *data* obtenida (Bender et al., 2021, p. 614).

Esta articulación entre la razón pública y la dimensión epistémica se torna especialmente exigente al considerar la estructura básica de la sociedad. Las instituciones de la estructura básica son el sujeto primario de la justicia porque sus efectos son profundos y están presentes desde el inicio de la vida de las personas (Rawls, 1993, p. 257). Ahora bien, la delegación continuada en algoritmos predictivos genera sesgos acumulativos que distorsionan silenciosamente las condiciones de trasfondo de la equidad estructural.

En esta escala sistémica, aunque cada micro-transacción individual pueda parecer justa tomada aisladamente, la interacción acumulativa en el tiempo degrada inevitablemente la

justicia de trasfondo si no existen reajustes institucionales de carácter estructural (Rawls, 1993, p. 261). De modo que la justificabilidad de la deferencia no puede evaluarse caso por caso, sino que requiere una consideración de los efectos sistémicos de la automatización decisional sobre la equidad de la estructura básica.

El principio de razón pública permite así establecer que la legitimidad de la deferencia algorítmica depende de la inteligibilidad normativa de sus premisas, la revisión deliberativa de los valores incorporados y el control de sus efectos sistémicos. Tales condiciones trasladan el análisis desde la precisión técnica hacia la justificabilidad intersubjetiva de la autoridad delegada. A partir de este estándar, la sección siguiente evalúa si los instrumentos contemporáneos de gobernanza algorítmica ofrecen garantías suficientes para satisfacer esas exigencias.

4. Límites de la gobernanza algorítmica ante la razón pública

Los marcos de gobernanza algorítmica desarrollados en la última década han articulado un conjunto de garantías orientadas a la regulación de los sistemas de inteligencia artificial en el sector público. El Reglamento de Inteligencia Artificial de la Unión Europea constituye el esfuerzo legislativo más ambicioso en esta dirección, combinando un enfoque basado en el riesgo con una lógica de seguridad de productos e imponiendo obligaciones de transparencia, documentación técnica y supervisión humana para los sistemas clasificados como de alto riesgo (Sousa e Silva, 2025, p. 4). Con todo, la capacidad de estos instrumentos para satisfacer el estándar de razón pública formulado en la sección precedente presenta limitaciones estructurales que conviene examinar con detenimiento.

Entre estas garantías institucionales, la transparencia formal, entendida como la disponibilidad de información sobre el funcionamiento del sistema, constituye la más frecuentemente invocada en la gobernanza algorítmica. Sin embargo, la mera revelación del "algoritmo" o la descripción de su lógica en lenguaje sencillo no equivale a una fundamentación normativa de las premisas decisorias. La regulación sectorial vigente, como la legislación del estado federado alemán de Schleswig-Holstein, exige la divulgación del algoritmo, pero admite excepciones amplias por confidencialidad pública o derechos de terceros que neutralizan el escrutinio cívico (Simbeck, 2022, p. 93). Además, la exigencia de transparencia se refiere al algoritmo y no al código fuente ni a los datos de entrenamiento, lo cual reduce la capacidad de los ciudadanos para evaluar críticamente los presupuestos incorporados en el sistema. De ahí que la transparencia formal resulte insuficiente si la información proporcionada no es inteligible

para los ciudadanos o si estos carecen de los recursos cognitivos para evaluar lo que se les revela.

Para hacer frente a esta insuficiencia de la transparencia formal, la Inteligencia Artificial Explicable (XAI) se propone como un avance técnico capaz de superar tales limitaciones. Las herramientas de XAI, tales como SHAP, LIME o la extracción de árboles de decisión, logran identificar con cierta eficacia los atributos o patrones que motivan la salida del sistema (Zednik y Verreault-Julien, 2026, p. 119). Ahora bien, estas herramientas presentan problemas de interpretabilidad residual, falta de robustez y dificultades de validación que comprometen su fiabilidad. Diferentes ejecuciones del mismo método pueden arrojar explicaciones distintas, y diferentes métodos, aplicados al mismo sistema, discrepan con regularidad (Zednik y Verreault-Julien, 2026, p. 122). Por tanto, las explicaciones *post hoc* de la XAI modelan el comportamiento del modelo sin revelar su funcionamiento real, y con frecuencia ofrecen recomendaciones de autodisciplina al usuario en lugar de razones públicas que permitan la contestación normativa (Hull, 2026, p. 20).

Más allá de estas falencias técnicas, la limitación más profunda de la XAI reside en su incapacidad estructural para suministrar razones normativas. Dado que los métodos de la XAI se diseñan para extraer información sobre el sistema mismo y no sobre su contexto sociotécnico, resultan incapaces de revelar los estándares éticos, legales o epistémicos externos que determinan si una decisión está justificada (Zednik y Verreault-Julien, 2026, p. 123). Aun si la ingeniería algorítmica lograse una transparencia técnica absoluta, el resultado seguiría huérfano de legitimidad si carece de razones normativas accesibles y justificables ante la ciudadanía. De este modo, la explicabilidad técnica constituye un insumo informativo valioso para la auditoría técnica y la evaluación del funcionamiento del sistema, pero no satisface el estándar de razón pública, el cual exige que las premisas decisorias puedan ser evaluadas en términos de justicia, reciprocidad y respeto a la autonomía.

Junto a las exigencias de explicabilidad, las métricas de equidad matemática representan otra herramienta central de la gobernanza algorítmica contemporánea. La investigación en esta área ha formulado múltiples estándares formales de equidad, entre los cuales figuran la paridad demográfica, la igualdad de oportunidades, el trato y el impacto dispar (Campione et al., 2026, p. 3). Sin embargo, un número considerable de estos criterios son matemáticamente incompatibles entre sí, lo cual impone elecciones normativas de fondo que no se expresan en el código, sino que permanecen ocultas bajo la apariencia de objetividad técnica. La selección de

una métrica determinada implica una ponderación entre valores en conflicto que corresponde a las partes interesadas y no a los científicos de datos (Henin y Le Métayer, 2021, p. 7). De modo que la reducción de la evaluación ética a un mero cumplimiento formal de estándares de equidad promueve la práctica del "lavado de equidad", una operación que proyecta una ilusión de legitimidad moral sin resolver las asimetrías de poder ni los daños estructurales subyacentes (Campione et al., 2026, p. 14).

Agravando esta insuficiencia de las métricas, emerge el problema irreductible de la singularidad individual. Los sistemas de aprendizaje automático reducen la complejidad humana a variables y vectores prefijados que, aun en configuraciones hiperpersonalizadas, solo ensamblan atributos predeterminados sin captar la irreductibilidad de cada caso particular (Campione et al., 2026, p. 7). Las clasificaciones estadísticas no son neutras ni naturales, sino que están moldeadas por paradigmas culturales dominantes cuya codificación consolida y reifica estructuras de dominación (Fricker, 2007, p. 14). Desde la perspectiva de la ética del cuidado, la pretendida imparcialidad del agente artificial no equivale a rectitud moral, ya que la falta de empatía e intersubjetividad constituye una falla ética en contextos donde la decisión afecta a personas concretas en situaciones de vulnerabilidad (Campione et al., 2026, p. 11).

Frente a las quiebras de estas garantías sustantivas, la contestabilidad ha sido postulada como la garantía procedimental complementaria de la transparencia y la equidad, entendida como la capacidad de los afectados para impugnar las decisiones algorítmicas y exigir una justificación de las mismas. Con todo, la contestabilidad *ex post* presenta una limitación estructural insuperable. La directiva algorítmica ya ha operado como una razón previamente establecida en la decisión administrativa, de suerte que el recurso gubernativo *a posteriori* no legitima la deferencia previa si la pretensión de autoridad del sistema no fue públicamente evaluada antes de su aplicación (Henin y Le Métayer, 2021, p. 8). La introducción de mecanismos como la "reprensión de IA" traslada toda la carga epistémica de impugnación al ciudadano una vez dictado el acto, con lo cual este debe demostrar la inadecuación de una decisión cuyas premisas no le son accesibles (Simbeck, 2022, p. 95). Tal inversión de la carga argumentativa resulta incompatible con las exigencias de la rendición de cuentas democrática.

En respaldo de este diagnóstico crítico, un estudio sobre la confiabilidad algorítmica muestra que las garantías teóricas rara vez se sostienen en la práctica. La distancia entre los valores éticos proyectados y los efectivamente materializados, junto con la falta de mecanismos de verificación en contextos reales, pone de manifiesto una brecha estructural entre las

formulaciones normativas y sus efectos concretos (Mehrotra et al., 2026, p. 32). La confiabilidad de la IA, lejos de ser una propiedad estática o un mero requisito de cumplimiento, emerge como un constructo disputado cuya definición depende de quienes tienen el poder de determinar sus parámetros (Mehrotra et al., 2026, p. 32:1).

Esta brecha de confiabilidad se agrava, de manera crítica, al considerar la naturaleza del error en los sistemas opacos. El desafío epistemológico decisivo no radica en ignorar cómo el sistema produce un resultado exitoso, sino en la imposibilidad de acceder a la naturaleza, mecanismo y fuente del error cuando este ocurre (Alvarado, 2026, p. 97). Las redes neuronales profundas fallan de manera "apacible", con desviaciones diminutas distribuidas a lo largo de millones de pesos, de tal modo que los errores se entregan con el mismo grado de confianza estadística que los resultados correctos (Alvarado, 2026, p. 102).

Por consiguiente, la fiabilidad frecuentista del sistema no garantiza que un resultado particular sea correcto, ni permite diagnosticar la causa del fallo cuando este se produce. Es decir, la deferencia práctica se sustenta en una presunción de fiabilidad que no puede ser verificada individualmente, lo cual transforma la aceptación del dictamen algorítmico en un acto de confianza ciega incompatible con la autonomía cívica. Ante tal insuficiencia, la sección siguiente desarrolla la desconfianza lúcida como respuesta epistemológico-política orientada a preservar la autonomía cívica y el control democrático de la autoridad algorítmica.

5. La desconfianza lúcida como actitud epistemológico-política

Frente al déficit de justificabilidad diagnosticado, la desconfianza lúcida se presenta como una actitud epistemológico-política normativamente adecuada y democráticamente exigible. Esta disposición no se fundamenta en un rechazo apriorístico de la tecnología ni en una forma de escepticismo radical que impida toda cooperación con sistemas computacionales. Se trata de una actitud derivada de la evaluación razonada de las condiciones bajo las cuales la deferencia podría estar justificada y de la constatación de que tales condiciones no se cumplen en la configuración actual de la gobernanza algorítmica. La desconfianza lúcida reconoce la posibilidad de que los sistemas algorítmicos desempeñen funciones valiosas en la toma de decisiones públicas, pero insiste en que dichas funciones deben estar sujetas a controles que garanticen que no socavan los principios fundamentales de la vida democrática.

El fundamento teórico de esta disposición reflexiva puede reconstruirse a partir del vínculo entre autoconfianza epistémica y autonomía. La autoconfianza ineludible del sujeto racional constituye la base de todo proceso reflexivo, y de ella se deriva la confianza en otros agentes

cuando se reconoce que estos encarnan las cualidades de escrupulosidad epistémica que el sujeto valora en sí mismo (Zagzebski, 2012, p. 99).

Sin embargo, cuando un sistema algorítmico opera bajo opacidad e invalida los canales de responsabilidad y escrutinio público, imposibilita que el sujeto efectúe un juicio reflexivo escrupuloso sobre la fuente (Zagzebski, 2012, p. 237). En tales circunstancias, la deferencia deja de ser un acto de autogobierno autónomo para convertirse en una subordinación heterónoma. La desconfianza lúcida se justifica, por tanto, como el ejercicio del principio de escrupulosidad que exige suspender la deferencia cuando las condiciones de justificación no se cumplen.

En su dimensión jurídico-política, la institucionalización de esta actitud implica el reconocimiento del derecho ciudadano a interrumpir la deferencia práctica hacia determinaciones algorítmicas cuyas premisas no son públicamente evaluables. Este derecho no es una concesión graciosa del legislador, sino una exigencia derivada del principio de libertad como no-dominación. La legitimidad del poder burocrático, en la tradición neo-republicana, se sustenta en que las decisiones institucionales permanezcan sujetas a la influencia y el control continuo de la ciudadanía (Frost, 2024, p. 783). Cuando la automatización desplaza la capacidad deliberativa del funcionario y reduce al ciudadano a un perfil estadístico, se produce una forma de dominación que no depende de la intención maliciosa de un agente, sino de la estructura misma del sistema decisional. Ante esta forma de dominación estructural, la interrupción de la deferencia opera como un mecanismo de autodefensa cívica.

Correlativamente, esta desconfianza conlleva una presunción de invalidez de los actos administrativos basados en sistemas cuya escrutabilidad no ha sido garantizada. Esta presunción no equivale a una anulación automática de toda decisión algorítmica, sino a una inversión de la carga argumentativa. En lugar de exigir al ciudadano que demuestre la inadecuación de la decisión, corresponde a la administración justificar públicamente por qué las premisas del sistema satisfacen el estándar de razón pública. La lógica de esta inversión se apoya en el principio de que, puesto que la autoridad reclama el derecho de reemplazar el juicio del sujeto sobre los méritos del caso (Raz, 1986, p. 58), debe ser posible identificar sus fallos autoritativos sin participar en un arduo argumento justificatorio (Raz, 1979, p. 51). Si la identificación de tales fallos resulta imposible debido a la opacidad del sistema, la presunción debe operar en favor del ciudadano.

Más allá de la impugnación de decisiones particulares, la vigilancia epistémica que exige la desconfianza lúcida abarca el cuestionamiento de la autoridad misma que se atribuye al sistema. La contestabilidad en sentido fuerte no se limita a la posibilidad de recurrir un acto administrativo concreto, sino que incluye la capacidad de impugnar los presupuestos estructurales del sistema, sus ponderaciones axiológicas, la calidad de sus datos de entrenamiento y la pertinencia de su función de optimización. Este nivel de contestación exige, a su vez, la construcción de capacidades cívicas específicas. La alfabetización algorítmica de los ciudadanos no es un lujo pedagógico, sino una condición necesaria para el ejercicio efectivo de la vigilancia democrática en un contexto de automatización creciente (Sousa e Silva, 2025, p. 10).

Junto a esta vigilancia ciudadana, la reconfiguración del principio de precaución epistémico-político constituye otra implicación relevante de la desconfianza lúcida. Ante sistemas cuya opacidad sobre la fuente y naturaleza del error resulta esencialmente insuperable (Alvarado, 2026, p. 102), la precaución no consiste en abstenerse de utilizar toda herramienta computacional, sino en subordinar la eficiencia predictiva a las exigencias de la razón pública. Ello significa que la decisión de delegar funciones de autoridad en un sistema algorítmico debe estar precedida de una evaluación pública de sus presupuestos y acompañada de mecanismos de revisión permanente. La eficiencia no puede operar como argumento autosuficiente para la delegación si la inteligibilidad de las premisas decisorias no está garantizada.

En el plano del diseño sociotécnico, la desconfianza lúcida se traduce en la exigencia de que los sistemas algorítmicos empleados en la decisión pública cumplan condiciones de resistibilidad y contestabilidad integradas en su propia arquitectura. La Protección Legal por Diseño postula que las garantías constitucionales deben traducirse en requerimientos técnicos que preserven la posibilidad de desobedecer al sistema y de impugnar sus determinaciones ante tribunales (Hildebrandt, 2015, p. 218). Si no se puede desobedecer a una ley, esta no es ley sino disciplina, y la misma lógica se aplica a las directivas algorítmicas que se imponen como hechos consumados sin posibilidad de contestación efectiva (Hildebrandt, 2015, p. 226).

Esta dimensión epistémica se articula orgánicamente con su dimensión política en la exigencia de preservar la naturaleza pública de la decisión administrativa. Las administraciones públicas funcionan como sitios de construcción democrática cuando se estructuran como comunidades de práctica integradas por redes dialógicas donde la pericia administrativa combina el conocimiento explícito con un conocimiento tácito adquirido mediante deliberación

e inmersión institucional (Frost, 2024, p. 787). La automatización algorítmica, al sustituir la deliberación comunitaria por la clasificación estocástica, destruye esta dinámica relacional y produce una "publicidad empobrecida" que desestabiliza el potencial democrático de la administración (Frost, 2024, p. 805). La desconfianza lúcida opera, en este contexto, como una salvaguarda institucional contra la descualificación profesional del funcionario público y la expropiación algorítmica de la articulación comunitaria del interés público.

Frente a este espacio de construcción comunitaria, la gobernanza algorítmica contemporánea erosiona la contestabilidad al sustituir el lenguaje de los derechos y las intenciones por un régimen estadístico de probabilidades inescrutables. La equidad aristotélica, como recurso de apelación e interpretación individualizada, permitía corregir la rigidez de la norma universal ante circunstancias particulares no previstas por el legislador (Hull, 2026, p. 5). Sin embargo, los sistemas algorítmicos codifican un compromiso fijo con lo que son los hechos administrativamente relevantes, y a menos que el sistema sea reentrenado, esos hechos permanecen inalterables (Hull, 2026, p. 22). La desconfianza lúcida reclama, frente a esta clausura, la reapertura del espacio deliberativo y la preservación del juicio prudencial como atributo irrenunciable de la decisión pública.

Lejos de constituir una postura definitiva, la desconfianza lúcida opera como un punto de partida para la construcción de criterios normativos capaces de distinguir entre formas de deferencia legítimas y formas que no lo son. Su institucionalización exige la creación de autoridades independientes de supervisión algorítmica, la implementación de auditorías *ex ante* de datos de entrenamiento, la obligación de justificación pública de los presupuestos del sistema antes de su despliegue y la garantía de mecanismos de revisión judicial efectiva. Cada uno de estos elementos responde a una condición del estándar de razón pública y contribuye a subordinar la eficiencia predictiva a la autonomía política de los ciudadanos.

6. Conclusiones

La atribución de autoridad epistémica a los sistemas algorítmicos en la toma de decisiones públicas exige un examen riguroso de sus condiciones de justificabilidad normativa. Cuando las estimaciones automatizadas operan como razones excluyentes para el decisor humano, la opacidad constitutiva de los modelos estocásticos impide evaluar de forma reflexiva su fundamento y conduce a una cesión heterónoma del juicio. La legitimidad de la deferencia práctica no puede sustentarse en la mera eficacia predictiva ni en la capacidad de procesar

grandes volúmenes de datos. Depende, por el contrario, de la adecuación de los presupuestos decisorios a las exigencias del razonamiento práctico intersubjetivo.

En el plano de la justificabilidad normativa, la articulación entre la teoría de la autoridad práctica y el principio de razón pública pone de manifiesto las insuficiencias del paradigma de gobernanza puramente técnico. Al distinguir las razones motivacionales, explicativas y normativas, se advierte que la explicabilidad técnica no equivale a la fundamentación pública exigida para el ejercicio de la coerción estatal. La formulación de un estándar de justificabilidad basado en la inteligibilidad normativa, la revisabilidad deliberativa de los valores codificados y el control de los efectos sistémicos sobre la justicia de trasfondo aporta una innovación conceptual relevante. Además, la caracterización de la desconfianza lúcida como postura epistemológico-política ofrece una alternativa tanto a la aceptación ciega como al rechazo tecnófobo y fundamenta el derecho cívico a interrumpir la deferencia automatizada.

Llevadas al plano de la gestión institucional, estas exigencias de razón pública imponen reestructurar los mecanismos normativos y procedimentales de la regulación algorítmica. Resulta indispensable pasar de la contestabilidad posterior y la transparencia formal a arquitecturas integradas de protección legal por diseño que subordinen la eficiencia predictiva a la deliberación pública. La administración debe asumir la carga de justificar la adecuación de los sistemas antes de su despliegue, invirtiendo la presunción habitual en favor del ciudadano. Del mismo modo, la implementación de auditorías públicas de los datos de entrenamiento, la evaluación deliberativa de las métricas de equidad y la capacitación cívica en alfabetización algorítmica constituyen condiciones necesarias para preservar el carácter público y prudencial de la función administrativa.

Más allá de la reestructuración administrativa, estas transformaciones en la infraestructura de la decisión pública abren nuevas líneas de investigación sobre la sostenibilidad democrática del Estado de derecho. Resulta prioritario examinar el impacto a largo plazo de la delegación algorítmica acumulativa sobre la equidad estructural y sobre la descualificación del juicio prudencial en la burocracia estatal. También corresponde analizar el diseño de instancias independientes de supervisión y la formulación de protocolos que permitan traducir patrones estocásticos de alta dimensionalidad a lenguajes normativos evaluables por la ciudadanía. Además, conviene ampliar el análisis hacia la relación entre la opacidad del error informático y la atribución de responsabilidad jurídica en contextos de gobernanza automatizada.

La preservación del espacio democrático exige recordar que la decisión pública es un acto moral relacional y no un problema de optimización estadística. La pretensión de sustituir la deliberación intersubjetiva por estimaciones probabilísticas amenaza con reducir al ciudadano a un perfil inerte y con despojar al poder burocrático de su función integradora. La institucionalización de una desconfianza lúcida no constituye un freno al progreso informático, sino una condición para que la tecnología permanezca subordinada al ideal republicano de la no dominación. En consecuencia, la legitimidad del ejercicio del poder depende de la capacidad colectiva para exigir razones comprensibles y compartibles frente a cualquier forma de autoridad, humana o artificial.

Bibliografía

Alvarado, R. (2026). "Challenges for Computational Reliabilism in AI and Other Computational Methods" en Durán, J. M. y Pozzi, G. (eds.), *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*, Synthese Library, vol. 527, 81-103. Cham: Springer.

Bender, E. M., Gebru, T., McMillan-Major, A. y Shmitchell, S. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" en *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610-623. Nueva York: ACM.

Campione, F., Lettieri, N. y Santucci, V. G. (2026). "Against Fair-Washing: A Critical Analysis of AI-Driven Decision Support Systems". *Minds and Machines*, 36 (21), 1-16.

Durán, J. M. (2026). "Beyond Transparency: Computational Reliabilism as an Externalist Epistemology of Algorithms" en Durán, J. M. y Pozzi, G. (eds.), *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*, Synthese Library, vol. 527, 55-79. Cham: Springer.

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Frost, N. (2024). "The Impoverished Publicness of Algorithmic Decision Making". *Oxford Journal of Legal Studies*, 44 (4), 780-807.

Hauswald, R. (2025). "Artificial Epistemic Authorities". *Social Epistemology*, 39 (6), 716-725.

Henin, C. y Le Métayer, D. (2021). "Beyond Explainability: Justifiability and Contestability of Algorithmic Decision Systems". *AI & Society: Knowledge, Culture and Communication*.

Hildebrandt, M. (2015). *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology*. Cheltenham: Edward Elgar Publishing.

Hull, G. (2026). “Erasing Agonism: How Algorithmic Governance Goes Further than Leviathan”. *Philosophy & Social Criticism*.

Mehrotra, S., Huang, J., Fu, X., Dobbe, R., Sánchez, C. I. y de Rijke, M. (2026). “Understanding AI Trustworthiness: A Scoping Review of AIES & FAccT Articles”. *Journal of Artificial Intelligence Research*, 85 (32), 1-31.

Ozer, A. L., Waggoner, P. D. y Kennedy, R. (2024). “The Paradox of Algorithms and Blame on Public Decision-Makers”. *Business and Politics*, 26 (1), 200-217.

Rawls, J. (1993). *Political Liberalism*. Nueva York: Columbia University Press.

Raz, J. (1979). *The Authority of Law: Essays on Law and Morality*. Oxford: Clarendon Press.

Raz, J. (1986). *The Morality of Freedom*. Oxford: Clarendon Press.

Simbeck, K. (2022). “FAccT-Check on AI Regulation: Systematic Evaluation of AI Regulation on the Example of the Legislation on the Use of AI in the Public Sector in the German Federal State of Schleswig-Holstein” en *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’22)*, 89-96. Nueva York: ACM.

Sousa e Silva, N. (2025). “The Artificial Intelligence Act: Critical Overview”. *Journal of Intellectual Property, Information Technology and E-Commerce Law (JIPITEC)*, 16 (1), 2-23.

Wang, F., Sohail, A., Wong, W.-K., Azim, Q. U. A., Farwa, S. y Sajad, M. (2023). “Artificial Intelligence and Stochastic Optimization Algorithms for the Chaotic Datasets”. *Fractals*, 31 (6), 2240175.

Zagzebski, L. T. (2012). *Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief*. Oxford: Oxford University Press.

Zednik, C. y Verreault-Julien, P. (2026). “Can XAI Justify?” en Durán, J. M. y Pozzi, G. (eds.), *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*, Synthese Library, vol. 527, 107-128. Cham: Springer.