Cómo citar: Páez, Andrés, Juan David Gutiérrez y Diana Acosta-Navas. "Hacia una inteligencia artificial centrada en los seres humanos: contribuciones de las ciencias sociales". *Revista de Estudios Sociales* 93: 3-18. https://doi.org/10.7440/res93.2025.01

Hacia una inteligencia artificial centrada en los seres humanos: contribuciones de las ciencias sociales*

Andrés Páez, Juan David Gutiérrez y Diana Acosta-Navas

Recibido: 6 de mayo de 2025 | Aceptado: 18 de junio de 2025 | Modificado: 7 de julio de 2025
https://doi.org/10.7440/res93.2025.01

Resumen | Esta introducción al *dossier* de la *Revista de Estudios Sociales,* dedicado a la relación entre las ciencias sociales y la inteligencia artificial (IA), parte de la premisa de que las ciencias sociales no solo tienen un papel fundamental en analizar y comprender las profundas transformaciones en la vida humana que ha implicado la irrupción de la IA, sino que también, y principalmente, tienen una función esencial en guiar el desarrollo futuro de la IA para que esté alineado con el bienestar humano y el desarrollo individual. Para asegurarnos de que el diseño, desarrollo, uso y evaluación de los sistemas de IA estén centrados en los seres humanos, necesitamos las herramientas conceptuales, los datos experimentales y la experiencia trabajando con diversas comunidades que pueden proporcionar las ciencias sociales. Una IA centrada en los humanos y alineada con sus intereses requiere de las ciencias sociales para existir. En esta introducción nos hemos centrado en tres áreas en las que consideramos que las ciencias sociales pueden cumplir ese papel de guía para el desarrollo futuro de la IA. Por una parte, consideramos que uno de los efectos más significativos de la IA en la vida humana es la manera en que está reconfigurando las relaciones interpersonales. La IA ha cambiado la forma en que las personas se comunican, forman relaciones, construyen su identidad, cooperan y enfrentan los desacuerdos. Por otra parte, examinamos el impacto de la IA en la igualdad social. Los modelos entrenados con datos históricos tienen el potencial de perpetuar las desigualdades y sesgos sociales reflejados en estos datos y producir resultados que ponen en desventaja a miembros de grupos marginados. Finalmente, nos interesa comprender cómo el desarrollo y uso de las diferentes aplicaciones de la IA influyen en los regímenes políticos, la política pública, la burocracia y el Estado.

Palabras clave | algoritmos de aprendizaje automático; discriminación algorítmica; IA y democracia; relaciones interpersonales; uso de la IA en las ciencias sociales

Toward Human-Centered Artificial Intelligence: Contributions from the Social Sciences

Abstract | This introduction to the *Revista de Estudios Sociales* dossier, which explores the intersection of artificial intelligence (AI) and the social sciences, is based on the idea that the social sciences are not only crucial for analyzing and understanding the major changes AI has brought to human life, but also—more fundamentally—for guiding the future development of AI in ways that support human well-being and individual growth. Ensuring that AI systems are designed, developed, used, and evaluated with

* Este artículo se escribió específicamente para este *dossier* y no contó con financiación. Los tres autores participaron en la escritura del artículo.

a human-centered focus requires the conceptual frameworks, empirical data, and community-based expertise that the social sciences offer. Human-centered AI—aligned with human values and interests—depends on these contributions to take shape. This introduction highlights three key areas where we believe the social sciences can provide meaningful guidance for the future direction of AI. First, we examine how AI is reshaping interpersonal relationships—transforming how people communicate, form bonds, construct identities, cooperate, and manage conflict. Second, we consider the impact of AI on social equality. Because models are often trained on historical data, they risk reproducing the inequalities and biases embedded in that data, which can lead to outcomes that disadvantage marginalized groups. Finally, we explore how the development and application of AI influence political regimes, policy-making, bureaucracy, and the role of the state.

Keywords | AI and democracy; algorithmic discrimination; interpersonal relationships; machine learning algorithms; use of AI in the social sciences

Rumo à inteligência artificial centrada no ser humano: contribuições das ciências sociais

Resumo | Esta introdução ao dossiê da *Revista de Estudios Sociales,* dedicado à relação entre as ciências sociais e a inteligência artificial (IA), parte da premissa de que as ciências sociais não apenas desempenham um papel fundamental na análise e compreensão das profundas transformações na vida humana que a irrupção da IA implicou, mas também, e sobretudo, exercem uma função essencial na orientação do desenvolvimento futuro da IA para que esteja alinhada ao bem-estar humano e ao desenvolvimento individual. Para garantir que o projeto, o desenvolvimento, o uso e a avaliação de sistemas de IA sejam centrados no ser humano, são necessários ferramentas conceituais, dados experimentais e conhecimento prático acumulado pelas ciências sociais no trabalho com diversas comunidades. Uma IA centrada nos seres humanos e alinhada a seus interesses depende das ciências sociais para existir. Nesta introdução, concentramo-nos em três áreas nas quais consideramos que as ciências sociais podem desempenhar esse papel como um guia para o desenvolvimento futuro da IA. Por um lado, acreditamos que um dos efeitos mais significativos da IA na vida humana é a maneira como ela está reconfigurando as relações interpessoais. A IA mudou a maneira como as pessoas se comunicam, formam relacionamentos, constroem sua identidade, cooperam e lidam com divergências. Por outro lado, examinamos o impacto da IA na igualdade social. Modelos treinados em dados históricos têm o potencial de perpetuar as desigualdades sociais e os vieses refletidos nesses dados e produzir resultados que prejudicam membros de grupos marginalizados. Por fim, interessa-nos compreender como o desenvolvimento e a utilização das diferentes aplicações da IA influenciam os regimes políticos, a política pública, a burocracia e o Estado.

Palavras-chave | algoritmos de aprendizado de máquina; discriminação algorítmica; IA e democracia; relações interpessoais; uso de IA nas ciências sociais

Introducción

Es difícil encontrar un aspecto de la vida humana que no se haya visto afectado por la irrupción de diversos sistemas computacionales basados en algoritmos de aprendizaje automático, popularmente conocidos como sistemas de inteligencia artificial (IA). Desde la administración de justicia hasta la automatización del trabajo, pasando por tareas tan diversas como la investigación científica y la creación artística, los algoritmos de aprendizaje automático han reconfigurado la relación de los seres humanos con la tecnología, han alterado profundamente las relaciones interpersonales, y han tenido repercusiones políticas, económicas y sociales que atraviesan todas las esferas de la vida humana.

Las ciencias sociales claramente ofrecen las herramientas críticas para analizar y entender estas profundas transformaciones. Sin embargo, parafraseando la famosa tesis 11 sobre Feuerbach[1], no basta con *interpretar* de diversas maneras los efectos de la IA en la sociedad; lo que importa es *transformar* el desarrollo futuro de la IA para que esté alineado con el bienestar humano y el desarrollo individual. No se trata simplemente de contribuir a la creación de marcos éticos y principios de gobernanza —que ciertamente son importantes— sino más bien de proporcionar las herramientas conceptuales, los datos experimentales y la experiencia trabajando con diversas comunidades para que el diseño, desarrollo, uso y evaluación de los sistemas de IA estén enfocados intrínsecamente en promover esos objetivos. Una IA centrada en los humanos y alineada con sus intereses requiere de las ciencias sociales para existir.

En la presentación de este *dossier* nos hemos enfocado en tres áreas en las que consideramos que las ciencias sociales pueden cumplir ese papel de guía para el desarrollo de la IA. En primer lugar, uno de los efectos más significativos de la IA en la vida humana reside en la manera en que está reconfigurando las relaciones interpersonales. Una comprensión adecuada de los cambios que genera la IA en estas relaciones puede servir para diseñar modelos y aplicaciones que combatan el aislamiento y preserven el valor agregado que proporciona la exposición a diversos puntos de vista en una sociedad. En segundo y tercer lugar, los efectos políticos de la IA, en particular sobre el orden social y la democracia, son cada día más evidentes. Por una parte, los modelos de IA suelen tener efectos sesgados y discriminatorios que reproducen las desigualdades existentes en la realidad social. Por la otra, la IA está siendo utilizada para generar discursos polarizados y extremistas que socavan los cimientos de las instituciones democráticas. En ambos frentes es necesario que las contribuciones de las ciencias sociales operen para encontrar mecanismos de neutralización de la polarización y soluciones al problema de las inequidades generadas por el uso negligente de los algoritmos. Existen herramientas tecnológicas que, si son utilizadas de manera correcta, pueden crear espacios de participación ciudadana más profundamente democráticos. Es claro, en consecuencia, que el trabajo de las ciencias sociales debe ir de la mano con el de las ciencias naturales y el de las ingenierías para crear un frente común que lidere el desarrollo responsable de la IA[2].

La IA en las relaciones interpersonales

Las ciencias sociales nos proporcionan herramientas conceptuales y metodológicas ideales para entender cómo los múltiples usos de la IA afectan las dinámicas interpersonales. La IA ha cambiado la forma en que las personas se comunican, forman relaciones, construyen su identidad, cooperan y enfrentan los desacuerdos. Nuestra forma de entender la amistad, el amor, la sexualidad, la creatividad, la comunicación, la confianza y la privacidad, entre muchos otros aspectos, se está viendo alterada por la irrupción de la IA. La psicología, la sociología y la antropología exploran justamente este tipo de efectos, apoyadas por las reflexiones políticas y éticas proporcionadas por la ciencia política y la filosofía. En este apartado nos ocuparemos de la contribución de las ciencias sociales a la comprensión del efecto disruptivo de la IA en las relaciones interpersonales.

1 Las "Tesis sobre Feuerbach" fueron escritas por Karl Marx en un cuaderno que nunca llegó a publicar. La tesis 11 reza: "Los filósofos se han limitado a interpretar el mundo de distintos modos; de lo que se trata es de transformarlo" (Marx 2014, 502).

2 La colaboración entre las ciencias sociales y naturales puede tomar múltiples formas. Por ejemplo, recientemente ha sido sugerido que el Consejo Internacional de la Ciencia (ISC, por sus siglas en inglés), que es la principal ONG mundial que integra las ciencias naturales y sociales, podría liderar dicha colaboración (ISC 2023).

Hoy en día, la IA está integrada a muchas de las herramientas tecnológicas que median las relaciones humanas: *chatbots*, aplicaciones de citas, redes sociales, entre otros. Estas herramientas tienen una enorme influencia sobre las dinámicas interpersonales al modificar los patrones de comunicación, la experiencia del amor, la amistad, la intimidad y la sexualidad (Delvin 2018; Gündüz 2017). La incertidumbre acerca de cómo enfrentar la irrupción de la tecnología en los espacios más íntimos del ser humano ha llevado a algunos a defender posiciones tan extremas como la prohibición de los robots sexuales (Richardson 2016). Sin embargo, existen perspectivas más mesuradas que se aproximan al asunto desde la salud sexual, la psicología, la economía y la ética de la sexualidad (Danaher y McArthur 2017; Döring *et al*. 2025).

Muchas compañías de tecnología hacen uso de perfilamientos psicológicos y de algoritmos de optimización de compromiso (*engagement*) con el fin de aumentar el tiempo de uso de las aplicaciones con fines comerciales. Uno de los efectos más negativos de este tipo de prácticas es la generación de burbujas epistémicas y cámaras de eco (Nguyen 2020) que aíslan a los individuos de opiniones diferentes a las suyas y refuerzan prejuicios y falsas creencias. Uno de los grandes retos es lograr que las personas tengan acceso a una gran diversidad de opiniones y puntos de vista, que las plataformas tecnológicas propicien lo que la socióloga Jane Jacobs (1961) llamaba *encuentros de acera*, propios de contextos diversos como lo son las grandes ciudades. Sin embargo, la diversidad de perspectivas puede afectar los patrones de uso de las aplicaciones, desincentivando a las compañías de tecnología a ampliar la oferta de contenido.

La IA, alimentada por las ciencias sociales, también puede ser usada para promover la confianza, la empatía, la justicia y la resolución de conflictos, por ejemplo, a través de los llamados *sistemas puente* (Acosta-Navas 2025b; Ovadya y Thorburn 2023). Estos sistemas se enfocan en las redes sociales, no como unos de distribución de contenido, sino como de distribución de atención. Al seleccionar los contenidos que capturan la atención de la mayoría de los usuarios es posible moldear los ambientes virtuales para evitar, por ejemplo, la polarización política. La idea es encontrar un terreno en común sobre el que se puedan construir desacuerdos racionales.

La psicología también puede contribuir al diseño de *chatbots* sociales y robots humanoides de gran utilidad en ámbitos como la educación y el cuidado de personas de la tercera edad. La relación entre humanos y robots ya es lo suficientemente común como para permitir un estudio sistemático. El desarrollo reciente de modelos de lenguaje de gran escala (LLM, por sus siglas en inglés) ha potenciado sistemas como los *chatbots* de compañía, los asistentes personales y los agentes de IA (Shin y Kim 2023). La forma en que los humanos antropomorfizamos y creamos lazos emocionales con las máquinas ha generado un fructífero campo de estudio en el que entran en juego conceptos como la confianza, la dependencia y el apego (Araujo *et al*. 2020; Lindgren y Holmström 2020; Páez 2021a). Mediante el estudio de las reacciones humanas ante la presencia de autómatas será posible ajustar su desarrollo para potenciar los beneficios que proveen.

La IA y el orden social

Una de las áreas más exploradas por la literatura en ética y ciencias sociales ha sido el impacto de la IA en la igualdad social. Estudios tempranos mostraron que los modelos entrenados con datos históricos tienen el potencial de perpetuar las desigualdades y sesgos sociales reflejados en estos datos y producir resultados que ponen en desventaja a miembros de grupos marginados. El estudio de los sesgos implícitos y explícitos hace parte fundamental de la psicología social y es necesario un trabajo conjunto entre los desarrolladores de estos sistemas y los científicos sociales para disminuir el riesgo de que estos sesgos sean reproducidos por las máquinas. Este impacto se ha visto en algoritmos

de contratación que favorecen candidatos hombres sobre mujeres (Chen 2023; Páez 2021b; Ramírez-Bustamante y Páez 2024); sistemas utilizados en la rama judicial para predecir la probabilidad de reincidencia en criminalidad cuya tasa de falsos positivos impacta desproporcionadamente a personas de raza negra (Angwin *et al*. 2016); o sistemas utilizados en programas de asistencia social (Eubanks 2018).

En muchos casos, sin embargo, los sesgos son solo un reflejo de las desigualdades existentes en la realidad, por ejemplo, en términos del acceso a la educación. Otros estudios han observado que los sesgos se pueden producir también por la definición de la variable objetivo, los *proxies* que se utilicen, e incluso por la manera en que los problemas son formalizados (Barocas y Selbst 2016; Barocas, Hardt y Narayanan 2023). En general, ha sido claro que las especificaciones del diseño de los modelos, los datos escogidos para entrenarlo y los algoritmos de IA pueden tener impacto de gran escala en el acceso a oportunidades y bienes sociales básicos por parte de diferentes grupos sociales. Una pregunta fundamental en el diseño de sistemas de IA es la relacionada con la definición de un algoritmo justo. El concepto de justicia puede ser visto desde diferentes perspectivas, y la filosofía ha contribuido a aclarar muchas de las discusiones en torno a los algoritmos justos (Barocas, Hardt y Narayanan 2023). Los modelos más recientes, conocidos como modelos fundacionales o LLM también exhiben sesgos, por ejemplo, en la generación de imágenes según estereotipos sociales (Bianchi *et al*. 2023). Un riesgo específico de los modelos fundacionales es la homogeneización de resultados: dado que son modelos sobre los cuales se construyen una gran cantidad de aplicaciones, existe el riesgo de que los sesgos y errores se transfieran a todas las aplicaciones que se construyen sobre ellos (Bommasani *et al*. 2021).

En la medida en que se utilicen para la generación y moderación de contenidos en línea, pueden también homogeneizar las perspectivas sobre temas de interés público, dando prioridad a visiones hegemónicas (Bender *et al*. 2021). En algunos casos, estos sesgos pueden ser deliberadamente introducidos por los desarrolladores, como es el caso de los modelos entrenados en China (como DeepSeek R1 y ErnieVG), que evitan cualquier mención o imagen relacionada con las protestas de Tiananmen Square (Lu 2025; Yang 2022). De manera similar, el intento de Google de incorporar lineamientos inclusivos en sus modelos llevó a que Gemini generara imágenes de personas de razas diversas cuando los usuarios le pedían imágenes de soldados nazis —claramente un error histórico— (Kleinman 2024; Raghavan 2024).

Tanto la IA predictiva como la IA generativa crean riesgos de reproducir y perpetuar injusticias históricas al limitar el acceso a recursos y oportunidades. También pueden reproducir y perpetuar estereotipos y perspectivas hegemónicas. Vale aclarar que estos riesgos pueden mitigarse en los procesos de diseño, entrenamiento y utilización de los sistemas de IA. Estos impactos pueden crear formas de desigualdad económica, social, y política. Por ello es de gran importancia que estos procesos estén informados por trabajo normativo, conceptual y empírico en las ciencias sociales, de tal manera que su impacto sea socialmente deseable (Floridi 2023).

Más allá de perpetuar y amplificar sesgos sociales, la IA puede generar otro tipo de desigualdad, que no depende de sus especificaciones de diseño. La adopción de la IA en este sentido puede llevar a una transformación de roles y perfiles laborales. En la medida en que esta es una tecnología que realiza tareas que normalmente requerirían inteligencia humana, genera el potencial de desplazamiento laboral para trabajadores de algunos sectores. Si bien otras tecnologías han reemplazado el trabajo humano, las tecnologías actuales han cambiado el tipo de trabajo que está expuesto a automatización, a saber, tareas cognitivas complejas en profesiones como periodismo, derecho, administración, e ingeniería de sistemas, entre otras (Spence 2022; Tyson y Zysman 2022).

No todos los trabajadores están expuestos por igual. Aquellos cuyos trabajos consisten exclusivamente de tareas que pueden ser satisfactoriamente realizadas por IA están expuestos a desplazamiento. Debido a que la IA puede realizar tareas con menor costo y mayor eficiencia, la posición de estos trabajadores en el mercado y el valor de su trabajo pueden verse negativamente afectados (Acemoglu y Restrepo 2018; Tyson y Zysman 2022). Aquellos cuyos trabajos solo pueden ser parcialmente realizados por IA pueden, por el contrario, utilizar la tecnología como una herramienta para complementar su propio trabajo y aumentar su eficiencia. Este último grupo puede beneficiarse de la IA. Al automatizar algunas tareas, la tecnología puede aumentar las capacidades de los trabajadores, generando mayor productividad y mayor valor, y ampliando el conjunto de tareas que puede realizar cada trabajador (Brynnolfsson 2022). El riesgo de utilizar IA para aumentar las capacidades productivas de un trabajador está en generar dependencia en la herramienta y, en el largo plazo, en la pérdida de habilidades. No es del todo claro cuáles son las circunstancias en las que es apropiado delegar tareas a estos sistemas debido a su tendencia al error. La respuesta frente a si los trabajadores serán o no reemplazados por la IA depende, no solo de la cantidad, sino también del tipo de tareas que estén expuestas a automatización (Winters y Latner 2025).

Un aspecto menos discutido de la relación de la IA con el trabajo es el surgimiento del llamado *trabajo fantasma*. El término fue acuñado en 2019 para referirse a los trabajadores humanos que laboran "tras bambalinas" para permitir el funcionamiento de los sistemas de IA (Gray y Suri 2019). Dos casos conocidos son los de los anotadores de datos en Kenia que hicieron el trabajo de *detoxificar* a ChatGPT (Perrigo 2023) y los moderadores de contenido que eliminan los contenidos tóxicos de las redes sociales (Roberts 2019). Estos trabajadores son personas cuyo trabajo es invisible para los usuarios de las tecnologías y que, en muchos casos, trabajan de manera informal a través de plataformas virtuales. Dado que esta categoría de trabajo es relativamente nueva y no del todo visible, existen riesgos asociados con la falta de entendimiento por parte del estado de las necesidades de estos trabajadores y la falta de regulación que garantice condiciones de trabajo dignas y seguras, y que protejan a los trabajadores de posibles daños psicológicos —como argumentan los trabajadores que demandaron a Meta en Ghana— (Hall y Wilmot 2025). En la medida en que la tecnología avanza y adquiere popularidad, crece la demanda por este tipo de trabajo.

En conjunto, estos cuatro efectos de la IA en el trabajo pueden generar una tendencia a ampliar las brechas de desigualdad social y económica. Unos cuantos sectores se pueden beneficiar de la introducción de herramientas que aumentan la eficiencia y la productividad, aumentando el valor de su trabajo. Otros sectores pueden estar expuestos al desplazamiento y a la pérdida de valor. Finalmente, la demanda por trabajo fantasma puede generar una tendencia a mayor informalidad laboral. Este tipo de desigualdad depende menos del diseño de las herramientas, y más de la utilización que se les dé en diferentes contextos. Depende también de la regulación en torno a su incorporación en espacios de trabajo y en torno a la economía de trabajos informales. De cualquier manera, esta es un área fértil para la investigación en ciencias sociales que puede contribuir a una mejor adopción de la tecnología que proteja las condiciones de igualdad y dignidad para los trabajadores más vulnerables.

Este problema forma parte de una categoría más amplia de retos que surgen cuando se analizan los impactos de la IA desde una perspectiva global. Dada la concentración de recursos en Estados Unidos y Europa, los países del sur global se encuentran subrepresentados tanto en el desarrollo de los modelos, como en su accesibilidad y las ventajas que generan, como en los lineamientos para su gobernanza. Organizaciones como el Instituto Equiano y el Movimiento Tierra Común han buscado traer perspectivas y herramientas de países africanos y latinoamericanos buscando que el diseño y desempeño de los modelos estén alineados con un conjunto más amplio de intereses y valores. El trabajo de Sabelo Mhlambi (2020) señala, por ejemplo, cómo incorporar la filosofía del Ubuntu para

diseñar una IA menos basada en la extracción de recursos y más orientada por las ideas de interdependencia y dignidad humana colectiva. Más recientemente, la iniciativa Global Dialogues empleó herramientas tecnológicas para realizar una consulta amplia respecto a las prioridades globales frente a la IA. En este *dossier* se busca contribuir a la expansión de perspectivas con el fin de obtener visiones de la IA desde una perspectiva latinoamericana.

Existe una tercera brecha de desigualdad que puede ser generada o ampliada por los sistemas de IA. A saber, una profunda desigualdad en la distribución del poder. Dada la escalabilidad de los sistemas de IA, estos tienen la posibilidad de generar impacto a gran escala en temas de interés público. Los modelos fundacionales, sobre los cuales se construyen múltiples aplicaciones, tienen un impacto aún mayor, puesto que los valores, sesgos y errores que se incorporen a estos modelos van a ser heredados por todas las aplicaciones que se construyan con base en ellos. Por ejemplo, cómo un modelo fundacional esté entrenado para identificar "contenido sexual inapropiado" puede afectar cómo se moderan los contenidos en redes sociales, afectando movimientos como el #MeToo e impactar desproporcionadamente contenidos de la comunidad LGBTIQ+. Esto puede tener impacto a la vez en sistemas de traducción, *chatbots* para atención médica y de salud mental, entre otros (Bender *et al*. 2021; Liang *et al*. 2023).

Dado que solo unos pocos desarrolladores tienen los recursos para entrenar modelos fundacionales, esto lleva a una concentración de poder entre los pocos actores privados que tienen los recursos para entrenar estos modelos. Por esta razón, las decisiones sobre el impacto social de estos sistemas recaerán sobre un grupo muy restringido de actores privados. Debido a que estas decisiones se toman en el contexto de empresas privadas, no existe ningún requisito de transparencia sobre estas decisiones ni mecanismos de rendición de cuentas. Si no se examina este problema, existe el riesgo de atentar contra el principio de igualdad política, según el cual cada ciudadano tiene igual poder de decisión sobre las cuestiones de interés público (Reich 2018; Saunders-Hastings 2022).

El impacto de los sistemas de IA sobre cuestiones de interés público genera preguntas acerca de la legitimidad de los actores privados que están a cargo de estas decisiones. Más en concreto, debemos plantear preguntas acerca de las estructuras, procesos, reglas y prácticas mediante las cuales se toman decisiones y se mantiene la rendición de cuentas dentro de las organizaciones que utilicen o desarrollen sistemas de IA. Estas preguntas cobrarán mayor importancia en la medida en que los sistemas de IA tengan la capacidad para realizar acciones complejas sin supervisión humana. Hoy en día los "agentes" de IA tienen capacidades muy limitadas en campos como la investigación y la programación. Sin embargo, los líderes de la industria se refieren a estos sistemas como la nueva frontera de desarrollo de la tecnología. Poniendo de lado cualquier especulación frente al desarrollo futuro de la tecnología, lo que parece ser claro es que sus capacidades para efectuar acciones (más allá de hacer predicciones y responder preguntas de sus usuarios) incrementa el poder de los desarrolladores para determinar el impacto social de estos sistemas.

Mantener a raya esta concentración de poder requiere de esfuerzos por parte de los mismos actores privados, organizaciones profesionales y regulación por parte de agentes estatales. Hoy en día hay diferentes regímenes regulatorios, por ejemplo, en Europa y en Estados Unidos, y la efectividad de la regulación parece estar atada a vientos políticos cambiantes. Por ello es importante que el sector privado se involucre en el esfuerzo de involucrar al público más amplio en decisiones de gran impacto. Algunas empresas han buscado crear procesos de consulta masiva, para obtener información sobre los valores e intereses de las partes afectadas por sus sistemas (Acosta-Navas 2025a; Perrigo 2024).

La IA, el Estado y la democracia

El estudio sobre las interacciones entre tecnologías, el Estado y la democracia se remonta por lo menos al siglo XIX, con pensadores clave para las ciencias sociales como Alexis de Tocqueville. Por ejemplo, Tocqueville exploró en su libro *La Democracia en América* ([1835] 2000), el rol de la prensa escrita para facilitar la asociación de las personas tanto para asuntos civiles como políticos. Otros ejemplos de los últimos dos siglos ilustran el interés histórico por comprender cómo el desarrollo y uso de diferentes tecnologías, entendidas como artefactos creados por humanos con fines prácticos, influyen en los regímenes políticos, la política, la burocracia y el Estado.

Es el caso de autores como Max Weber, que discutió el rol de los expedientes y archivos en el establecimiento de las burocracias estatales (1984); de Lewis Mumford, que documentó la influencia recíproca, a lo largo de la historia, entre tecnologías y organizaciones sociopolíticas (1992); de Herbert Marcuse, que exploró cómo la automatización industrial modificaba la relación Estado-ciudadano, posibilitando el Estado de bienestar y aumentado el control sobre la libertad individual (1993); de Langdon Winner, que argumentó que las tecnologías no son instrumentos neutrales sino que pueden tener dimensiones políticas cuando "se convierten en un medio para alcanzar un determinado fin dentro de una comunidad" o cuando "parecen requerir, o ser altamente compatibles con, determinados tipos de relaciones políticas" (1980, 123); y, más recientemente, de Shoshana Zuboff (2019), que planteó el concepto *capitalismo de vigilancia*, según el cual las empresas se valen de tecnologías digitales —que hacen parte de nuestro día a día— para recolectar masivamente información, predecir conductas e intentar moldearlas, y cuyas dinámicas concentran el poder y debilitan las democracias.

Es pertinente mencionar la larga trayectoria en el estudio de las tecnologías en las ciencias sociales porque puede informar las investigaciones sobre las implicaciones de los sistemas de IA respecto del Estado y la democracia. Las herramientas de IA tienen características que las distinguen, particularmente sus diversos niveles de autonomía para alcanzar objetivos, pero comparten características con otras herramientas digitales tales como la complejidad, la operación a partir de datos, la apertura y vulnerabilidad a ciberataques (European Commission 2019). Por ejemplo, las líneas de investigación sobre las interacciones entre la IA, el Estado y la democracia, tienen vasos vinculantes con los estudios sobre tecnologías como el internet, las redes sociales, las tecnologías *blockchain*, los sistemas algorítmicos basados en reglas utilizados para automatizar, entre otros.

En relación con la democracia, las ciencias sociales han estudiado los efectos del uso de herramientas de IA en contextos electorales y respecto de otras esferas públicas esenciales para los procesos democráticos. En cuanto a lo primero, la literatura ha explorado el uso de IA generativa para desinformar a través de la creación de contenido ultrafalsificado (*deepfakes*) en forma de imágenes, audio y video que pueden afectar la integridad electoral (Régis *et al*. 2025).

Las ciencias sociales también han estudiado cómo el discurso político está siendo filtrado, controlado y manipulado en redes sociales a partir de algoritmos de IA que persiguen fines ajenos al bienestar social, que fomentan la polarización y que dificultan la deliberación democrática (Sunstein 2017). Además, las ciencias sociales también han explorado cómo las empresas y los Estados, secretamente, están empleando la IA con fines de vigilancia, control social y para restringir el libre flujo de la información, lo que pone en peligro derechos y libertades individuales y colectivas (Pasquale 2015).

Por otra parte, la literatura también ha comenzado a explorar los potenciales efectos —positivos y negativos— de herramientas basadas en IA en el contexto de "plazas públicas digitales" como el internet. Por ejemplo, Goldberg *et al*. (2024) han examinado cómo herramientas basadas en LLM pueden contribuir a mejorar las conversaciones que tienen lugar en estas esferas. Por una parte, el uso de estas herramientas puede aportar positivamente, facilitando diálogos delicados, moderando el discurso tóxico, recompensando el discurso que tiende puentes y sintetizando los resultados de deliberaciones a gran escala. Simultáneamente, el uso de estas herramientas también plantea riesgos como exacerbar sesgos y patrones discriminatorios contra poblaciones subrepresentadas, fomentar la homogenización del discurso y la manipulación de la opinión.

En materia de gestión y administración pública, la literatura ha investigado la aplicación de herramientas de IA por parte de entidades públicas. En la última década, ha crecido significativamente la aplicación de herramientas de IA por parte de entidades de la rama ejecutiva, judicial y legislativa y con ello el interés por entender cómo son integradas en las organizaciones y procesos públicos (Medaglia y Tangi 2022; Misuraca, Van Noordt y Boukli 2020; Sousa *et al*. 2019; Wirtz y Müller 2019).

Este proceso ha sido estudiado en América Latina y el Caribe, por ejemplo, a partir de la creación de bases de datos que documentan dichas aplicaciones y de reportes que caracterizan los hallazgos del mapeo de sistemas automatizados de toma de decisiones (GobLab UAI 2022; Gutiérrez y Muñoz-Cadena 2023; Sistemas de Algoritmos Públicos 2025). También se ha estudiado el grado en que el uso de estos sistemas se corresponde con el principio de transparencia algorítmica y con las iniciativas de Estado Abierto (Gutiérrez y Castellanos-Sánchez 2023; Hermosilla, Garrido y Loewe 2020; Hermosilla y Lapostol 2022; Lapostol, Garrido y Hermosilla 2023; Valderrama, Hermosilla y Garrido 2023).

Además del estudio de las contribuciones de estas tecnologías a la generación de valor público, también ha crecido el interés por entender sus implicaciones negativas, especialmente las afectaciones a los derechos humanos tales como el derecho a la no discriminación, el debido proceso y a la privacidad (Flórez Rojas y Vargas Leal 2020; Gutiérrez 2024a; Risse 2019).

En el caso de África, por ejemplo, Effoduh, Akpudo y Kong (2024) han explorado cómo las debilidades en la gobernanza de los datos para el uso de herramientas de IA por parte de los Estados han generado nuevas vulnerabilidades para la población en términos de vulneración de protección de datos personales, viralización de desinformación en contextos electorales, y discriminación y nuevas formas de segregación racial a través de medios digitales. Recientes estudios sobre el continente africano también han identificado casos de "uso dañino de IA" asociado a armas autónomas y vigilancia estatal a partir de sistemas de reconocimiento facial (Akpudo *et al*. 2024).

Finalmente, en materia de política pública, además de una generosa literatura que ha estudiado la emergencia de políticas públicas y estrategias nacionales de IA alrededor del mundo, también se ha examinado el auge de otras herramientas de gobernanza de IA como las guías y marcos éticos (Corrêa *et al*. 2023), y la regulación de la IA alrededor del mundo (Gutiérrez 2024b). Además, tanto en Europa como en América Latina, la literatura ha explorado el rol de las herramientas en las diferentes etapas del ciclo de las políticas públicas (Gutiérrez y Muñoz-Cadena 2025; Valle-Cruz *et al*. 2020).

Usos de la IA en la investigación en ciencias sociales

La IA también es una herramienta de investigación de gran utilidad en el estudio de las dinámicas interpersonales y el discurso en línea sobre una gran variedad de temas. La IA permite examinar grandes volúmenes de datos, tales como redes sociales, encuestas, registros estatales, entre otros, para encontrar tendencias y correlaciones demasiado complejas o sutiles para los métodos estadísticos tradicionales. Por ejemplo, los modelos de procesamiento de lenguaje natural (NLP) y de computación afectiva son entrenados con la información que tenemos acerca de estas dinámicas para ayudarnos a entender las emociones, intenciones y comportamientos humanos en escenarios dialógicos y sociales. Las herramientas de análisis de sentimientos entrenadas utilizando datos psicométricos ayudan a las empresas a entender las emociones de sus clientes en tiempo real. Es importante señalar, sin embargo, que estas herramientas tienen limitaciones. Por ejemplo, la herramienta falla cuando la emoción se expresa de manera sarcástica o usando *slang*, cuando hay polisemia o cuando la persona tiene emociones mezcladas (Wankhade, Sehhara Rao y Kulkarni 2022).

La IA también permite simular cómo interactúan los individuos bajo diferentes condiciones con la ayuda de los modelos multiagente (Anthis *et al*. 2025; Park *et al*. 2024). Estos modelos permiten predecir comportamientos grupales o tendencias culturales al estudiar sistemas sociales complejos. También son útiles entendiendo cómo se disemina la desinformación y cómo se genera la cooperación entre grupos. Este tipo de modelos también pueden combinarse con datos históricos para hacer predicciones de diversos tipos, incluyendo predicciones electorales y tendencias económicas y comerciales.

En la investigación cualitativa de las interacciones humanas, la IA puede utilizarse para detectar cuestionarios que contengan sesgos y para mejorar su redacción (Krägeloh, Alyami y Medvedev 2023). Igualmente, la IA puede ser útil en la generación de cuestionarios adaptativos en tiempo real, en los que las preguntas que se generan dependen de las respuestas anteriores. La IA también puede adaptar los cuestionarios para que correspondan a determinados grupos demográficos o etarios, y permite analizar cómo diferentes formulaciones u ordenamiento de las preguntas pueden afectar los resultados. El uso de agentes conversacionales, además, puede facilitar la aplicación de los cuestionarios, especialmente, en personas que tienen dificultad leyendo y entendiendo un gran número de preguntas. Finalmente, utilizando técnicas de NLP es posible codificar y resumir las respuestas a las preguntas abiertas en un cuestionario de manera más eficiente que los programas estadísticos tradicionales, y hacer un análisis de las emociones y actitudes detectadas en los textos. Estas características permiten el uso de técnicas cualitativas en una mayor escala con menor tiempo de procesamiento.

Uno de los mayores riesgos en el uso de herramientas cualitativas es el comportamiento fraudulento y las respuestas de baja calidad. En ambos frentes la IA puede ser de gran ayuda. El riesgo de fraude ocurre principalmente cuando se utilizan empresas de recolección virtual de información usando incentivos económicos. En muchos casos las respuestas son proporcionadas por *bots* que usan libretos prediseñados. La IA puede detectar patrones poco naturales en las respuestas, como respuestas repetidas o tiempos muy similares de respuesta, e incluso puede analizar los metadatos para detectar que múltiples respuestas bajo identidades diferentes provienen de la misma IP (Lebrun *et al*. 2024). Las respuestas de baja calidad ocurren cuando los participantes resuelven de una manera precipitada los cuestionarios o escogen siempre la misma opción para terminar rápidamente la tarea. También puede ocurrir que, en el caso de las preguntas abiertas, copien y peguen los mismos textos. Incluso es posible detectar si el contenido de estas preguntas proviene de internet o si ha sido copiado de otras secciones del mismo cuestionario (Badarovski 2024).

Presentación de los artículos

Los cinco artículos reunidos en este *dossier* desarrollan varios de los temas tratados en esta Introducción. Los primeros tres utilizan técnicas de inteligencia artificial para analizar diferentes aspectos de la vida social: el discurso de género, las concepciones del bienestar y las gramáticas discursivas sobre riesgos y seguridad asociadas a la IA. Los otros dos artículos hacen reflexiones más teóricas sobre dos aspectos centrales del uso de la IA: la idea de que los algoritmos deben ser "justos" y evitar la discriminación algorítmica, y los problemas que surgen en el uso de algoritmos de decisión automatizada en el derecho.

En el artículo de Laura Manrique Gómez, Tony Montes y Rubén Manrique, "Gender Semantics and Historical Feminisms: An Interdisciplinary Approach through Natural Language Processing", se emplean técnicas de IA aplicadas a un corpus de periódicos históricos de varios países latinoamericanos, para identificar y analizar las concepciones de género implícitas en el uso del término "mujeres" a lo largo del siglo XIX. Los autores demuestran cómo se puede utilizar el procesamiento de lenguaje natural para revelar tensiones implícitas en la cultura y los cambios en normas sociales, dejando claro cómo la IA puede proporcionar herramientas enriquecedoras para la investigación en ciencias sociales.

De manera similar, en el artículo "Mapping Poverty and Well-Being Through Natural Language Processing", de Guberney Muñetón-Santa, Carlos Andrés Pérez-Aguirre y Juan Rafael Orozco-Arroyave, se utilizan técnicas de modelamiento temático para identificar dimensiones de pobreza y bienestar a partir del lenguaje natural expresado por personas en el contexto de entrevistas. Con estas herramientas, los autores extraen dimensiones significativas como empleo, salud y necesidades básicas, entre otras, con el fin de informar el diseño de políticas públicas orientadas a la generación de capacidades. El estudio demuestra cómo el procesamiento de lenguaje natural permite dar un lugar más central a las voces de la gente (en sus propias palabras y no mediadas por encuestas) para el diseño de indicadores multidimensionales significativos y representativos de los intereses y necesidades de la población.

En el artículo de Mónica A. Ulloa Ruiz y Guillem Bas Graells, "La securitización de la inteligencia artificial: un análisis de sus impulsores y sus consecuencias", se realiza un análisis crítico de veinticinco actos discursivos de agencias gubernamentales, organizaciones técnico-científicas y medios de comunicación de Estados Unidos, para identificar gramáticas discursivas sobre riesgos y seguridad asociada a la IA. Los autores argumentan que la IA está siendo configurada por estos agentes políticos en Estados Unidos como una cuestión de seguridad que amerita un tratamiento extraordinario. En el artículo se discuten las tensiones que surgen en relación con diferentes lógicas del proceso de securitización (por ejemplo, amenaza vs. riesgo) y se identifican las diferencias discursivas y tensiones narrativas que emergen entre tipos de actores políticos (por ejemplo, militarización vs. regulación, nacional vs. global). Por último, los autores aportan sus reflexiones sobre las implicaciones del proceso de securitización para la gobernanza de la IA.

Por su parte, en el artículo "Evitando la trampa del formalismo: evaluación crítica y selección de métricas de equidad estadística en algoritmos públicos", de Alberto Coddou Mc Manus, Mariana Germán Ortiz y Reinel Tabares Soto, encontramos una reflexión teórica sobre el concepto de justicia algorítmica. Los autores muestran que no basta con encontrar definiciones técnicas precisas de qué es un algoritmo justo, es decir, un algoritmo que no arroje resultados discriminatorios, sino que también se debe tener en cuenta que en diferentes contextos sociales se deben privilegiar diferentes formas de medir ese tipo de justicia. Los autores introducen pautas normativas que indican qué métricas de equidad deberían priorizarse en función del contexto de uso del algoritmo.

Finalmente, en el artículo de María Lorena Flórez Rojas, "*Algocracy* in the Judiciary: Challenging Trust in the System", se investiga, a través de estudios de caso, cómo el uso de sistemas de toma automatizada de decisiones en la administración de justicia puede afectar tanto la legitimidad como la confianza percibida por la ciudadanía respecto de los órganos judiciales. El estudio se centra en el estudio de la *algocracia* en la justicia, entendida como la delegación de la toma de decisiones judiciales a sistemas algorítmicos. La autora argumenta que los estudios de caso de México, Australia, Alemania, Colombia, los Países Bajos y Dinamarca ilustran cómo las deficiencias en el uso de este tipo de herramientas debilitan la percepción de la confiabilidad de los órganos judiciales y, por lo tanto, ponen en peligro su legitimidad institucional.

Referencias

1. Acemoglu, Daron y Pascual Restrepo. 2018. "Artificial Intelligence, Automation, and Work". En *The Economics of Artificial Intelligence: An Agenda*, editado por Ajay Agrawal, Joshua Gans y Avi Goldfarb, 197-236. Chicago: University of Chicago Press.
2. Acosta-Navas, Diana. 2025a. "On Foundations and Foundation Models: What AI and Philanthropy can Learn from One Another". En *The Routledge Handbook of Artificial Intelligence and Philanthropy*, editado por Giuseppe Ugazio y Milos Maricic, 393-407. Nueva York; Londres: Routledge.
3. Acosta-Navas, Diana. 2025b. "In Moderation: Automation in the Digital Public Sphere". *Journal of Business Ethics*. https://doi.org/10.1007/s10551-024-05912-8
4. Akpudo, Ugochukwu Ejike, Jake Okechukwu Effoduh, Jude Dzevela Kong y Yongsheng Gao. 2024. "Unveiling AI Concerns for Sub-Saharan Africa and its Vulnerable Groups". En *2024 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, editado por Sameerchand Pudaruth y Kingsley Ogudo, 45-55. https://doi.org/10.59200/ICONIC.2024.007
5. Angwin, Julia, Jeff Larson, Surya Mattu y Lauren Kirchner. 2016. "Machine Bias". *ProPublica*, 23 de mayo. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
6. Anthis, Jacy Reese, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, *et al*. 2025. "LLM Social Simulations Are a Promising Research Method". *arXiv*. https://arxiv.org/abs/2504.02234
7. Araujo, Theo, Natali Helberger, Sanne Kruikemeier y Claes H. De Vreese. 2020. "In AI We Trust? Perceptions about Automated Decision-Making by Artificial Intelligence". *AI & Society* 35: 611-623. https://doi.org/10.1007/s00146-019-00931-w
8. Badarovski, Marjan. 2024. "The Role of AI in Enhancing Survey Data Quality: How AI Can Help Detect Fraudulent Responses and Improve Panel Management". *Lifepanel Blog*, 27 de septiembre. https://lifepanel.eu/blog/the-role-of-ai-in-enhancing-survey-data-quality-how-ai-can-help-detect-fraudulent-responses-and-improve-panel-management/
9. Barocas, Solon, y Andrew D. Selbst. 2016. "Big Data's Disparate Impact". *California Law Review* 104 (3): 671-732. http://dx.doi.org/10.2139/ssrn.2477899
10. Barocas, Solon, Moritz Hardt y Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, MA: MIT Press.
11. Bender, Emily M., Timnit Gebru, Angelina McMillan-Major y Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?". En *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623. 3 al 10 de marzo, virtual, Canadá. https://doi.org/10.1145/3442188.3445922
12. Bianchi, Federico, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, *et al*. 2023. "Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale". En *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493-1504. 12 al 15 de junio, Chicago, Estados Unidos. https://doi.org/10.1145/3593013.3594095
13. Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, *et al*. 2021. "On the Opportunities and Risks of Foundation Models". *ArXiv.* https://doi.org/10.48550/arXiv.2108.07258
14. Brynjolfsson, Erik. 2022. "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence". *Dædalus* 151 (2): 272-287. https://www.amacad.org/publication/daedalus/turing-trap-promise-peril-human-artificial-intelligence
15. Chen, Zhisheng. 2023. "Ethics and Discrimination in Artificial Intelligence-Enabled Recruitment Practices". *Humanities and Social Sciences Communications* 10: 1-12. https://doi.org/10.1057/s41599-023-02079-x
16. Corrêa, Nicholas Kluge, Camila Galvão, James William Santos, Carolina Del Pino, Edson Pontes Pinto, Camila Barbosa, *et al*. 2023. "Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance". *Patterns* 4 (10): 1-14. https://doi.org/10.1016/j.patter.2023.100857

17. Danaher, John, y Neil McArthur, eds. 2017. *Robot Sex: Social and Ethical Implications*. Cambridge, MA: MIT Press.
18. Delvin, Kate. 2018. *Turned On: Science, Sex and Robots*. Londres: Bloomsbury.
19. Döring, Nicola, Thuy Dung Le, Laura M. Vowels, Matthew J. Vowels y Tiffany L. Marcantonio. 2025. "The Impact of Artificial Intelligence on Human Sexuality: A Five-Year Literature Review 2020-2024". *Current Sexual Health Reports* 17: 1-39. https://doi.org/10.1007/s11930-024-00397-y
20. Effoduh, Jake Okechukwu, Ugochukwu Ejike Akpudo y Jude Dzevela Kong. 2024. "Toward a Trustworthy and Inclusive Data Governance Policy for the Use of Artificial Intelligence in Africa". *Data & Policy* 6: 1-14. https://doi.org/10.1017/dap.2024.26
21. Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. Nueva York: St. Martin's Press.
22. European Commission. 2019. "Liability for Artificial Intelligence and other Emerging Technologies". European Commission [edición digital]. https://data.europa.eu/doi/10.2838/573689
23. Flórez Rojas, María Lorena y Juliana Vargas Leal. 2020. "El impacto de herramientas de inteligencia artificial: Un análisis en el sector público en Colombia". En *Inteligencia artificial en América Latina y el Caribe. Ética, gobernanza y políticas*, editado por Carolina Aguerre, 206-238. Buenos Aires: CETyS; Universidad de San Andrés.
24. Floridi, Luciano. 2023. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford: Oxford University Press.
25. GobLab UAI. 2022. "Repositorio de algoritmos públicos de Chile. Primer informe de estado de uso de algoritmos en el sector público". Santiago de Chile: Universidad Adolfo Ibáñez (UAI). https://goblab.uai.cl/wp-content/uploads/2022/02/Primer-Informe-Repositorio-Algoritmos-Publicos-en-Chile.pdf
26. Goldberg, Beth, Diana Acosta-Navas, Michiel Bakker, Ian Beacock, Matt Botvinick, Prateek Buch, *et al*. 2024. "AI and the Future of Digital Public Squares". *arXiv*. https://arxiv.org/abs/2412.09988
27. Gray, Mary L. y Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Nueva York: Harper Business.
28. Gutiérrez, Juan David. 2024a. "Critical Appraisal of Large Language Models in Judicial Decision-Making". En *Handbook on Public Policy and Artificial Intelligence*, editado por Regine Paul, Emma Carmel y Jennifer Cobbe, 323-338. Londres: Edward Elgar Publishing.
29. Gutiérrez, Juan David. 2024b. *Consultation Paper on AI Regulation: Emerging Approaches across the World*. París: Unesco. https://unesdoc.unesco.org/ark:/48223/pf0000390979
30. Gutiérrez, Juan David y Michelle Castellanos-Sánchez. 2023. "Transparencia algorítmica y Estado abierto en Colombia". *Reflexión Política* 25 (52): 6-21. https://doi.org/10.29375/01240781.4789
31. Gutiérrez, Juan David y Sarah Muñoz-Cadena. 2023. "Adopción de sistemas de decisión automatizada en el sector público: Cartografía de 113 sistemas en Colombia". *GIGAPP Estudios Working Papers* 10 (270): 365-395. https://www.gigapp.org/ewp/index.php/GIGAPP-EWP/article/view/329
32. Gutiérrez, Juan David y Sarah Muñoz-Cadena. 2025. "Artificial Intelligence in Latin America's Public Policy Cycles". En *Handbook of Public Policy in Latin America*, editado por Leonardo Secchi y César N. Cruz-Rubio, 403-420. Londres: Edward Elgar Publishing.
33. Gündüz, Uğur. 2017. "The Effect of Social Media on Identity Construction". *Mediterranean Journal of Social Sciences* 8 (5): 85-92. http://archive.sciendo.com/MJSS/mjss.2017.8.issue-5/mjss-2017-0026/mjss-2017-0026.pdf
34. Hall, Rachel y Claire Wilmot. 2025. "Meta Faces Ghana Lawsuits over Impact of Extreme Content on Moderators". *The Guardian*, 27 de abril. https://www.theguardian.com/technology/2025/apr/27/meta-faces-ghana-lawsuits-over-impact-of-extreme-content-on-moderators
35. Hermosilla, María Paz, Romina Garrido y Daniel Loewe. 2020. "Transparencia y responsabilidad algorítmica para la inteligencia artificial". Santiago de Chile: Gob_Lab UAI/Escuela de Gobierno Universidad Alfonso Ibañez. https://goblab.uai.cl/wp-content/uploads/2020/04/Paper-Transparencia-GobLab.pdf
36. Hermosilla, María Paz y José Pablo Lapostol. 2022. "The Limits of Algorithmic Transparency". En *Survey on the Use of Information and Communication Technologies in the Brazilian Public Sector: ICT Electronic Government 2021*, editado por el Núcleo de Informação e Coordenação do Ponto BR, 289-295. Brasil: Comitê Gestor da Internet no Brasil.
37. ISC (International Science Council). 2023. *A Framework for Evaluating Rapidly Developing Digital and Related Technologies: AI, Large Language Models and Beyond*. París: ISC. https://council.science/publications/framework-digital-technologies/
38. Jacobs, Jane. 1961. *The Death and Life of Great American Cities*. Nueva York: Random House.
39. Kleinman, Zoe. 2024. "Why Google's 'Woke' AI Problem Won't Be an Easy Fix". *BBC News*, 24 de febrero. https://www.bbc.com/news/technology-68412620
40. Krägeloh, Christian U., Mohsen M. Alyami y Oleg N. Medvedev. 2023. "AI in Questionnaire Creation: Guidelines Illustrated in AI Acceptability Instrument Development". En *International Handbook of Behavioral Health Assessment*, 1-23. Cham: Springer.

41. Lapostol, José Pablo, Romina Garrido y María Paz Hermosilla. 2023. "Algorithmic Transparency from the South: Examining the State of Algorithmic Transparency in Chile's Public Administration Algorithms". En *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency,* 227-235. 12 al 15 de junio, Chicago, Estados Unidos. https://doi.org/10.1145/3593013.3593991
42. Lebrun, Benjamin, Sharon Temtsin, Andrew Vonasch y Christoph Bartneck. 2024. "Detecting the Corruption of Online Questionnaires by Artificial Intelligence". *Frontiers in Robotics and AI* 10: 1-25. https://doi.org/10.3389/frobt.2023.1277635
43. Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, *et al*. 2023. "Holistic Evaluation of Language Models". *Transactions on Machine Learning Research*. https://openreview.net/pdf?id=iO4LZibEqW
44. Lindgren, Simon y Jonny Holmström. 2020. "A Social Science Perspective on Artificial Intelligence: Building Blocks for a Research Agenda". *Journal of Digital Social Research* 2 (3): 1-15. https://doi.org/10.33621/jdsr.v2i3.65
45. Lu, Donna. 2025. "We Tried Out DeepSeek. It Worked Well, until we Asked It about Tiananmen Square and Taiwan". *The Guardian*, 28 de enero. https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan
46. Marcuse, Herbert. 1993. *El hombre unidimensional*. Barcelona: Planeta DeAgostini.
47. Marx, Karl. 2014. "Tesis sobre Feuerbach". En *Karl Marx, La ideología alemana*, 499-502. Madrid: Akal.
48. Medaglia, Rony y Luca Tangi. 2022. "The Adoption of Artificial Intelligence in the Public Sector in Europe: Drivers, Features, and Impacts". En *ICEGOV '22: Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance,* 10-18. 4 al 7 de octubre, Guimarães, Portugal. https://doi.org/10.1145/3560107.3560110
49. Mhlambi, Sabelo. 2020. "Q&A: Sabelo Mhlambi on What AI Can Learn from Ubuntu Ethics". *People + AI Research Blog*, *Medium,* 6 de mayo. https://medium.com/people-ai-research/q-a-sabelo-mhlambi-on-what-ai-can-learn-from-ubuntu-ethics-4012a53ec2a6
50. Misuraca, Gianluca, Colin van Noordt y Anys Boukli. 2020. "The Use of AI in Public Services: Results from a Preliminary Mapping across the EU". En *ICEGOV '20: Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, 90-99. 23 al 25 de septiembre, Atenas, Grecia. https://doi.org/10.1145/3428502.3428513
51. Mumford, Lewis. 1992. *Técnica y civilización*. Madrid: Alianza Editorial.
52. Nguyen, C. Thi. 2020. "Echo Chambers and Epistemic Bubbles". *Episteme* 17 (2): 141-161. https://doi.org/10.1017/epi.2018.32
53. Ovadya, Aviv y Luke Thorburn. 2023. "Bridging Systems: Open Problems for Countering Destructive Divisiveness across Ranking, Recommenders, and Governance". *arXiv.* https://arxiv.org/abs/2301.09976
54. Páez, Andrés, ed. 2021a. "Robot Mindreading and the Problem of Trust". *AISB Convention 2021: Communication and Conversation*, 140-143. Londres: AISB.
55. Páez, Andrés. 2021b. "Negligent Algorithmic Discrimination". *Law and Contemporary Problems* 84 (3): 19-33. https://scholarship.law.duke.edu/lcp/vol84/iss3/3
56. Park, Joon Sung, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, *et al*. 2024. "Generative Agent Simulations of 1,000 People". *arXiv*. https://arxiv.org/abs/2411.10109
57. Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.
58. Perrigo, Billy. 2023. "Exclusive: OpenAI Used Kenyan Workers on Less than $2 per Hour to Make ChatGPT Less Toxic". *Time*, 18 de enero. https://time.com/6247678/openai-chatgpt-kenya-workers/
59. Perrigo, Billy. 2024. "Inside OpenAI's Plan to Make AI more 'Democratic'". *Time*, 5 de febrero. https://time.com/6684266/openai-democracy-artificial-intelligence/
60. Raghavan, Prabhakar. 2024. "Gemini Image Generation Got It Wrong. We'll Do Better". *Google The Keyword*, 23 de febrero. https://blog.google/products/gemini/gemini-image-generation-issue/
61. Ramírez-Bustamante, Natalia y Andrés Páez. 2024. "Análisis jurídico de la discriminación algorítmica en los procesos de selección laboral". En *Derecho, poder y datos. Aproximaciones críticas al derecho y las nuevas tecnologías*, editado por René Urueña Hernández y Natalia Ángel-Cabo, 203-230. Bogotá: Ediciones Uniandes.
62. Régis, Catherine, Florian Martin-Bariteau, Okechukwu Jake Effoduh, Juan David Gutiérrez, Gina Neff, Carlos Affonso Souza, *et al*. 2025. "AI in the Ballot Box: Four Actions to Safeguard Election Integrity and Uphold Democracy". *IVADO*. https://www.uottawa.ca/research-innovation/sites/g/files/bhrskd326/files/2025-02/gpb-ai_ai_and_elections.pdf
63. Reich, Rob. 2018. *Just Giving: Why Philanthropy Is Failing Democracy and How It Can Do Better*. Princeton: Princeton University Press.
64. Richardson, Kathleen. 2016. "The Asymmetrical 'Relationship': Parallels Between Prostitution and the Development of Sex Robots". *ACM SIGCAS Computers and Society* 45 (3): 290-293. https://doi.org/10.1145/2874239.2874281
65. Risse, Mathias. 2019. "Human Rights and Artificial Intelligence". *Human Rights Quarterly* 41 (1): 1-16. https://doi.org/10.1353/hrq.2019.0000
66. Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven: Yale University Press.

67. Saunders-Hastings, Emma. 2022. *Private Virtues, Public Vices: Philanthropy and Democratic Equality*. Chicago: University of Chicago Press.
68. Shin, Minkyu, Jin Kim y Jiwoong Shin. 2023. "The Adoption and Efficacy of Large Language Models: Evidence from Consumer Complaints in the Financial Industry". *arXiv*. https://arxiv.org/abs/2311.16466
69. Sistemas de Algoritmos Públicos. 2025. "Informe de los Repositorios 1.0". Bogotá: Escuela de Gobierno / Universidad de los Andes. https://sistemaspublicos.tech/wp-content/uploads/Informe-1-Repositorios-Proyecto-SAP-032025-PLATAFORMA.pdf
70. Spence, Andrew Michael. 2022. "Automation, Augmentation, Value Creation & the Distribution of Income & Wealth". *Dædalus* 151 (2): 244-255. https://www.amacad.org/publication/daedalus/automation-augmentation-value-creation-distribution-income-wealth
71. Sunstein, Cass R. 2017. *#Republic: Divided Democracy in the Age of Social Media*. Princeton: Princeton University Press.
72. Tyson, Laura D. y John Zysman. 2022. "Automation, AI & Work". *Dædalus* 151 (2): 256-271. https://www.amacad.org/publication/daedalus/automation-ai-work
73. Tocqueville, Alexis de. (1835) 2000. *Democracy in America*. Chicago: University of Chicago Press.
74. Valderrama, Matías, María Paz Hermosilla y Romina Garrido. 2023. "State of the Evidence: Algorithmic Transparency". *Open Government Partnership*, 24 de mayo. https://www.opengovpartnership.org/documents/state-of-the-evidence-algorithmic-transparency/
75. Valle-Cruz, David, J. Ignacio Criado, Rodrigo Sandoval-Almazán y Edgar A. Ruvalcaba-Gomez. 2020. "Assessing the Public Policy-Cycle Framework in the Age of Artificial Intelligence: From Agenda-Setting to Policy Evaluation". *Government Information Quarterly* 37 (4): 101509. https://doi.org/10.1016/j.giq.2020.101509
76. Wankhade, Mayur, Annavarapu Chandra Sekhara Rao y Chaitanya Kulkarni. 2022. "A Survey on Sentiment Analysis Methods, Applications, and Challenges". *Artificial Intelligence Review* 55 (7): 5731-5780. https://doi.org/10.1007/s10462-022-10144-1
77. Weber, Max. 1984. *Economía y sociedad*. Ciudad de México: FCE.
78. Winner, Langdon. 1980. "Do Artifacts Have Politics?". *Dædalus* 109 (1): 121-136. https://www.jstor.org/stable/20024652?seq=1
79. Winters, Jutta y Jonathan Latner. 2025. "Does Automation Replace Experts or Augment Expertise? The Answer Is Yes". *IAB-Forum*, 9 de enero. https://www.iab-forum.de/en/does-automation-replace-experts-or-augment-expertise-the-answer-is-yes/
80. Wirtz, Bernd W. y Wilhelm M. Müller. 2019. "An Integrated Artificial Intelligence Framework for Public Management". *Public Management Review* 21 (7): 1076-1100. https://doi.org/10.1080/14719037.2018.1549268
81. Yang, Zeyi. 2022. "There's no Tiananmen Square in the New Chinese Image-Making AI". *MIT Technology Review*, 14 de septiembre. https://www.technologyreview.com/2022/09/14/1059481/baidu-chinese-image-ai-tiananmen/
82. Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism*. Nueva York: Routledge.

Andrés Páez

Ph.D. en Filosofía por The City University of New York, Estados Unidos. Profesor titular del Departamento de Filosofía e investigador del Centro de Investigación y Formación en Inteligencia Artificial (CinfonIA) de la Universidad de los Andes, Colombia. Sus investigaciones recientes en inteligencia artificial se centran en los problemas éticos y epistemológicos de la opacidad algorítmica. Publicaciones recientes: "Understanding with Toy Surrogate Models in Machine Learning", *Minds and Machines* 34 (45): 1-26, 2024, https://doi.org/10.1007/s11023-024-09700-1; y "Explainability of Algorithms", en *A Companion to Digital Ethics*, editado por Luciano Floridi y Mariarosaria Taddeo, (Oxford: John Wiley & Sons, Ltd, 2025). https://orcid.org/0000-0002-4602-7490 | apaez@uniandes.edu.co

Juan David Gutiérrez

Ph.D. en Política Pública por University of Oxford, Reino Unido. Profesor asociado de la Escuela de Gobierno de la Universidad de los Andes, Colombia. Investiga y enseña sobre política pública, inteligencia artificial, competencia y regulación, y gobernanza de los recursos naturales. Publicaciones recientes: "Artificial Intelligence in Latin America's Public Policy Cycles" (en coautoría), en *Handbook of Public Policy in Latin America*, editado por Leonardo Secchi y César N. Cruz-Rubio, 403-420 (Londres: Edward Elgar Publishing, 2025); y "Critical Appraisal of Large Language Models in Judicial Decision-Making", *Handbook on Public Policy and Artificial Intelligence*, editado por Regine Paul, Emma Carmel y Jennifer Cobbe, 323-338 (Londres: Edward Elgar Publishing, 2024). https://orcid.org/0000-0002-7783-4850 | juagutie@uniandes.edu.co

Diana Acosta-Navas

Ph.D. en Filosofía por Harvard University, Estados Unidos. Profesora asistente de Ética en Loyola University Chicago, Estados Unidos. Su investigación examina el papel de las plataformas digitales en el fomento de un debate público positivo y las implicaciones éticas de las tecnologías emergentes en la esfera pública digital. También explora los derechos humanos en situaciones de posconflicto, centrándose en la responsabilidad moral de las empresas para apoyar los procesos de consolidación de la paz. Publicaciones recientes: "On Foundations and Foundation Models: What AI and Philanthropy can Learn from One Another", en *The Routledge Handbook of Artificial Intelligence and Philanthropy*, editado por Giuseppe Ugazio y Milos Maricic, 393-407 (Nueva York; Londres: Routledge, 2025); e "In Moderation: Automation in the Digital Public Sphere", *Journal of Business Ethics*, 2025, https://doi.org/10.1007/s10551-024-05912-8. https://orcid.org/0000-0002-5351-5820 | dacostanavas@luc.edu