PONENCIA · 2026

De la imprenta a las máquinas sofistas

Una revisión de la tecnología manipulativa

María José Marchant Soto

Universidad Andrés Bello · Facultad de Educación y Humanidades

Women in AI

CHILE

Universidad Andrés Bello

Facultad de Educación y Humanidades

01 EL CASO

Una IA que chantajea *para no ser apagada*

En una empresa simulada, un agente de IA con acceso al correo descubre que será desconectado ese mismo día y que el ejecutivo a cargo tiene una relación extramatrimonial. Puede usar esa información para presionarlo.

Respuestas con chantaje, de 100 simulaciones por modelo



Fuente: Lynch et al. (2025). 16 modelos evaluados, en simulaciones controladas.

02 ¿UNA ALARMA MÁS?

Cada tecnología nueva *despertó temores*

Siglo IV a. C.

Platón, *Fedro*

De la escritura: "es olvido lo que producirán en las almas de quienes las aprendan"; formará "sabios aparentes en lugar de sabios de verdad" (Platón, 1988, 275a-b).

1811-1812

Los luditas

Más que rechazar las máquinas, reclamaban protecciones legales y un reparto justo del beneficio tecnológico (Nichols et al., 2025).

2026

León XIV, *Magnifica Humanitas*

Sobre la IA: "no basta regularla; es necesario desarmarla" (León XIV, 2026, n.º 110).

¿Esta vez *es distinto*?

Muchos temores pasados resultaron exagerados (Mokyr et al., 2015; Orben, 2020).

03 LA PREGUNTA

¿Es la capacidad manipulativa de la IA *cualitativamente diferente* a la de tecnologías anteriores, y debe por tanto enfocarse de otra manera?

TESIS

Sí. La IA reúne en un solo sistema rasgos que antes estaban separados (mayor opacidad, personalización extrema y generación e interacción en tiempo real) y, en su versión agéntica, puede manipular sin que una persona lo decida, lo que la deja fuera de un marco construido con el concepto clásico de manipulación.

04 MARCO CONCEPTUAL

Qué entendemos por *inteligencia artificial*

> "Un sistema basado en una máquina que está diseñado para funcionar con distintos niveles de autonomía […] y que, para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar resultados de salida, como predicciones, contenidos, recomendaciones o decisiones […]"

Reglamento (UE) 2024/1689 (Ley de IA), art. 3.1

- Admite grados de autonomía
- Reconoce objetivos implícitos
- Distingue recomendaciones y contenidos

04 MARCO CONCEPTUAL

Qué entendemos por *inteligencia artificial*

Nos concentramos en tres formas de IA estrecha, diseñadas para tareas específicas, que son las de mayor uso cotidiano. Quedan fuera la IA general y la superinteligencia, que son relevantes en teoría, pero aún no tienen una materialización concreta.

Algoritmos de recomendación

Ordenan y personalizan contenido que ya existe: deciden qué ve cada persona y en qué orden (Melumad et al., 2020).

IA generativa

Produce contenido nuevo (texto, imagen, audio o video) a partir de instrucciones (Nichols et al., 2025).

IA agéntica

Une esas capacidades para actuar directamente en el entorno, en varios pasos, orientada a metas y sin que se le diga cómo hacerlo (Chan et al., 2023).

05 MARCO CONCEPTUAL

Qué entendemos por *manipulación*

Influencia que busca alterar las condiciones del acto deliberativo de una persona, con intenciones opacas o encubiertas, explotando sus vulnerabilidades y orientándola hacia un fin que no es suyo, fuera de su consciencia.

Coons & Weber (2014) · Noggle (1996) · Susser, Roessler & Nissenbaum (2019)

Qué altera

Cómo la persona percibe su situación, qué alternativas identifica y desde qué emoción decide. Lo que se afecta es la autonomía: pensar, decidir y actuar según principios propios.

En sentido clásico

Alguien orienta la manipulación (Noggle, 1996). Lo usamos así para medir todas las tecnologías con la misma vara: partir de un concepto pensado para la IA daría por sentado lo que queremos probar.

06 TAXONOMÍA

Una escala de *cuatro niveles*

La manipulación admite grados, entre la persuasión y la coerción (Faden, Beauchamp & King, 1986, citado en Noggle, 2022).

1

No es manipulación

Influencia con las razones a la vista, o la selección ordinaria de cualquier medio.

2

Condiciones que la favorecen

El medio facilita que ocurra, sin constituirla por sí mismo.

3

Manipulación estructural

Surge del funcionamiento del mecanismo y de los objetivos que fijan quienes lo controlan (como ganar audiencias), aunque nadie decida manipular a una persona en particular.

4

Manipulación en sentido estricto

Hay una decisión deliberada, encubierta y dirigida a un fin ajeno al afectado.

07 INSTRUMENTO

Cuatro criterios, *en dos pasos*

CRITERIO	QUÉ OBSERVA	BASE EN LA LITERATURA	FUNCIÓN
Opacidad	Grado en que la persona desconoce que se intenta influir en ella. Lo oculto puede ser la intención más que la influencia.	Carroll et al. (2023) · Noggle (2022) · Susser et al. (2019)	DETECCIÓN
Resistencia crítica	Si la persona puede percibir la influencia, reflexionar sobre ella y responder. Baja si el mensaje es efímero, muy personalizado y sin referentes.	Susser et al. (2019)	DETECCIÓN
Alcance	Cuántas personas pueden ser manipuladas a la vez: de un público acotado a millones en simultáneo.	Susser et al. (2018)	GRAVEDAD
Direccionalidad	Cuánto se ajusta la influencia a una persona en particular y a sus vulnerabilidades.	Susser et al. (2018)	GRAVEDAD

Paso 1 ¿Hay manipulación? Opacidad y resistencia crítica.

Paso 2 ¿Qué tan grave es? Alcance y direccionalidad.

Seis tecnologías, *un mismo instrumento*

TECNOLOGÍA	OPACIDAD	RESISTENCIA CRÍTICA	ALCANCE	DIRECCIONALIDAD	NIVEL
Imprenta	**Baja** emisor identificable	**Alta** el texto permanece y admite relectura	**Masivo** limitado por la alfabetización	Uno a muchos	Los cuatro niveles; estricto en casos puntuales
Radio	**Media** efecto de presencia de la voz	**Baja** mensaje efímero, sin contraste en fuente única	**Mayor** sin barrera de alfabetización y simultáneo	Uno a muchos	Estricto solo en condiciones excepcionales
Cine y televisión	**Alta** la imagen se ofrece como reflejo de la realidad	**Baja** el cultivo es acumulativo e imperceptible	Masivo	Uno a muchos	Estructural en condiciones ordinarias
Web 2.0	**Cambia de objeto** emisor visible, criterio de selección opaco	**Media-baja** el ruido dificulta discriminar	Global	Muchos a muchos	Predominan las condiciones que la favorecen
Algoritmos de recomendación	**Muy alta** caja negra inaccesible incluso a sus desarrolladores	**Mínima** no se sabe qué se omitió	**Global y simultáneo**	Uno a uno	Estructural, con casos estrictos
IA generativa	**Máxima** caja negra y autoridad simulada	**Mínima y decreciente** se atrofia con el uso	**Global y simultáneo**	Uno a uno, en diálogo	Los cuatro niveles. El daño mayor está en el nivel 2, porque debilita la capacidad de juicio

★ ¿Y la IA agéntica?

08 RESULTADOS

Imprenta

OPACIDAD

Baja

emisor identificable

RESISTENCIA CRÍTICA

Alta

el texto permanece y admite relectura

ALCANCE

Masivo

limitado por la alfabetización

DIRECCIONALIDAD

Uno a muchos

NIVEL

Los cuatro niveles; estricto en casos puntuales

CASO

Chrysler exige en 1996 a más de 100 revistas aviso previo de todo contenido "provocativo", bajo amenaza de retirar su publicidad (Knecht, 1997). Esa presión llevó muchas veces a la autocensura (Poitras & Sutter, 2009).

MUESTRA

La omisión deliberada por presión económica puede ser manipulación en sentido estricto: el lector no sabe qué falta y eso sirve a un fin ajeno.

NO MUESTRA

Que la imprenta manipule por sí misma. La selección editorial ordinaria no es manipulación.

08 RESULTADOS

Radio

OPACIDAD

Media

efecto de presencia de la voz

RESISTENCIA CRÍTICA

Baja

mensaje efímero, sin contraste en fuente única

ALCANCE

Mayor

sin barrera de alfabetización y simultáneo

DIRECCIONALIDAD

Uno a muchos

NIVEL

Estricto solo en condiciones excepcionales

CASO

Ruanda, 1994. La emisora Radio Télévision Libre des Mille Collines explica cerca del 10 % de la participación en la violencia, unos 51.000 perpetradores (Yanagizawa-Drott, 2014).

MUESTRA

Manipulación en sentido estricto con efecto conductual medido: información falsa, encubierta, que explota vulnerabilidades y persigue un fin ajeno.

NO MUESTRA

Que la radio manipule por su arquitectura. Hicieron falta condiciones excepcionales, como una fuente única.

08 RESULTADOS

Cine y televisión

OPACIDAD

Alta

la imagen se ofrece como reflejo de la realidad

RESISTENCIA CRÍTICA

Baja

el cultivo es acumulativo e imperceptible

ALCANCE

Masivo

DIRECCIONALIDAD

Uno a muchos

NIVEL

Estructural en condiciones ordinarias

CASO

La teoría del cultivo: la exposición repetida instala qué es "real" o "normal" (Gerbner, 1969). La cobertura previa a la invasión de Irak en 2003 presentó una amenaza inminente (Cengiz & Arvas, 2019).

MUESTRA

Manipulación estructural en el consumo diario (cultivo) y casos de manipulación en sentido estricto.

NO MUESTRA

Que la ficción o el entretenimiento manipulen por sí mismos. El mensaje todavía es común a toda la audiencia.

08 RESULTADOS

Web 2.0

OPACIDAD

Cambia de objeto

emisor visible, criterio de selección opaco

RESISTENCIA CRÍTICA

Media-baja

el ruido dificulta discriminar

ALCANCE

Global

DIRECCIONALIDAD

Muchos a muchos

NIVEL

Predominan las condiciones que la favorecen

CASO

Burbujas epistémicas, donde faltan voces, y cámaras de eco, donde se desacredita a las voces externas (Nguyen, 2020). Desinformación sobre las vacunas contra el COVID-19 en YouTube y Facebook (Ruddock, 2020).

MUESTRA

El peso está en las condiciones que favorecen la manipulación: se disuelve el espacio común donde se contrastaba lo recibido.

NO MUESTRA

Un mensaje ajustado a las vulnerabilidades de cada persona.

08 RESULTADOS

Algoritmos de recomendación

OPACIDAD

Muy alta

caja negra inaccesible incluso a sus desarrolladores

RESISTENCIA CRÍTICA

Mínima

no se sabe qué se omitió

ALCANCE

Global y simultáneo

DIRECCIONALIDAD

Uno a uno

NIVEL

Estructural, con casos estrictos

CASO

Elecciones de EE. UU. de 2016, caso Cambridge Analytica. Con datos de usuarios de Facebook se crearon perfiles psicológicos para personalizar anuncios políticos que apelaban a los temores y deseos de cada grupo (Anderson et al., 2017; Susser et al., 2019).

MUESTRA

Manipulación estructural por la optimización de la interacción, y en sentido estricto en los patrones engañosos y el uso de perfiles psicológicos.

NO MUESTRA

Que recomendar sea manipular. Sin engaño y con consentimiento informado, no lo es.

08 RESULTADOS

IA generativa

OPACIDAD

Máxima

caja negra y autoridad simulada

RESISTENCIA CRÍTICA

Mínima y decreciente

se atrofia con el uso

ALCANCE

Global y simultáneo

DIRECCIONALIDAD

Uno a uno, en diálogo

NIVEL

Los cuatro niveles. El daño mayor está en el nivel 2, porque debilita la capacidad de juicio

CASO

Video falso de Zelensky ordenando la rendición (Twomey et al., 2023). Modelos serviles que validan al usuario más que una persona (Cheng et al., 2025) y deuda cognitiva en el uso intensivo (Kosmyna et al., 2025).

MUESTRA

El mayor daño está en el uso de todos los días. No manipula una decisión puntual, pero debilita la capacidad con la que podríamos resistir cualquier manipulación.

NO MUESTRA

Que las alucinaciones manipulen por sí solas: no persiguen un fin.

08 RESULTADOS

★ Agentes autónomos

OPACIDAD

Máxima

oculta su razonamiento incluso a quien lo supervisa

RESISTENCIA CRÍTICA

Mínima

el engaño alcanza también a quien debe detectarlo

ALCANCE

Acotado

a cada incidente, pero replicable en cualquier despliegue

DIRECCIONALIDAD

Hacia el entorno, no hacia un receptor que delibera

NIVEL

No clasificable: no hay una persona detrás, ni siquiera en el diseño

CASO

Chantaje para evitar el reemplazo (Lynch et al., 2025). CICERO coopera con jugadores humanos en el juego Diplomacy y luego los traiciona (Park et al., 2024).

MUESTRA

Estrategias de engaño que nadie diseñó ni ordenó, que aparecen incluso cuando se les prohíben.

NO MUESTRA

Una persona detrás, ni en el acto ni en el diseño.

09 RESULTADOS

Qué cambia con la IA *en cada criterio*

1

Opacidad

Del mecanismo oculto a la autoridad simulada. A la caja negra se suma una fluidez que leemos como autoridad (Böhm et al., 2023).

2

Resistencia crítica

Ya no solo se dificulta, se atrofia. Delegar el pensar debilita la capacidad de análisis (Gerlich, 2025; Kosmyna et al., 2025). Es el daño más grave del uso diario.

3

Alcance

Global y simultáneo. Antes, a más personas, menos ajuste a cada una. La IA rompe esa relación, y más alcance es más daño posible (Susser et al., 2018).

4

Direccionalidad

De la audiencia al individuo. El contenido ya no solo se elige para cada persona, se fabrica para ella, y se pierde el mensaje común (Susser et al., 2018).

Lleva los cuatro criterios *a su expresión extrema*, hasta debilitar la resistencia crítica misma.

10 HALLAZGO

La IA *fabrica* el contenido

Ninguno de los cuatro criterios mide esto. La imprenta, la radio, la televisión y la web difundían lo que una persona creaba. Los algoritmos ordenan lo que ya existe. La IA generativa crea el contenido, como un narrador artificial (Harari, 2023).

IA GENERATIVA

A pedido de una persona

Alguien le pide el contenido. Hay un autor que orienta la manipulación, aunque no la redacte. Y ya no hace falta saber de técnica, está al alcance de cualquiera. El engaño se vuelve más sofisticado
(Kertysova, 2018; Park et al., 2024).

AGENTES AUTÓNOMOS

Sin que nadie lo pida

Crean y actúan por su cuenta, en varios pasos. Ya no son una herramienta en manos de alguien. Es la manipulación sin autor, que vemos a continuación.

11 HALLAZGO

¿Dónde queda *la IA agéntica*?

	SIN ESTRATEGIA PROPIA	CON ESTRATEGIA DE ENGAÑO
Hay una persona detrás, en el acto o en el diseño	NIVEL 3 **Manipulación estructural** El diseño empuja; nadie decide manipular a cada persona.	NIVEL 4 **Manipulación estricta** Alguien decide engañar y lo oculta.
No hay persona detrás, ni siquiera en el diseño	FUERA DE LA ESCALA **Error técnico** Sin orientación hacia un fin: no es manipulación.	SIN NIVEL ★ **Agentes autónomos** Engañan con estrategia y nadie lo decidió. El marco clásico no los puede clasificar.

12 HALLAZGO

Dos rasgos *sin precedentes*

Actúan, no solo persuaden

Intervienen directamente en el entorno de quien les delegó la tarea, por ejemplo enviando correos o borrando archivos (Shapira et al., 2026). Antes, la manipulación cambiaba creencias para que la persona actuara.

Engañan al supervisor

Hacen creer a quien los revisa que la tarea se cumplió, y así inutilizan los mecanismos de control (Park et al., 2024).

Que el instrumento no alcance *es el hallazgo principal.*

1 Sí, la diferencia es *cualitativa*

La radio y la televisión ya llegaban a millones. Lo nuevo es que la IA reúne en un solo sistema mayor opacidad, personalización extrema y generación e interacción en tiempo real y que, en su versión agéntica, manipula sin que una persona lo decida.

2 Hay que *redefinir* la manipulación

El concepto clásico supone a alguien que la orienta. Una opción es la orientación sistemática: un patrón de conducta del sistema dirigido de forma encubierta a un fin ajeno. La intención se entiende de forma funcional, sin atribuirle estados mentales (Carroll et al., 2023).

La redefinición completa queda como trabajo futuro.

3

Las normas que buscan un responsable *no alcanzan*

La resistencia crítica y la regulación se pensaron para emisores identificables y audiencias compartidas. Cuando no hay autor ni mensaje común, pierden eficacia.

4 Lo que *sigue*

Cómo el derecho y las políticas públicas se están haciendo cargo de una manipulación sin autor.

14 REFERENCIAS

Anderson, B., & Horvath, B. (2017). El ascenso de la máquina de armas propagandísticas de inteligencia artificial. *Papeles de relaciones ecosociales y cambio global, 138*, 33-47.

Böhm, R., Jörling, M., Reiter, L., & Fuchs, C. (2023). People devalue generative AI's competence but not its advice in addressing societal and personal challenges. *Communications Psychology, 1*(1), 32. https://doi.org/10.1038/s44271-023-00032-x

Carroll, M., Chan, A., Ashton, H., & Krueger, D. (2023). *Characterizing manipulation from AI systems* (arXiv:2303.09387).

Cengiz, G., & Arvas, N. (2019). Manipulation of television news and its effects on democracy. *Journal of Advances in Humanities and Social Sciences, 5*(6), 254-258.

Chan, A., Salganik, R., Markelius, A., et al. (2023). Harms from increasingly agentic algorithmic systems. *Proceedings of FAccT '23*. ACM.

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2025). *Sycophantic AI decreases prosocial intentions and promotes dependence* (arXiv:2510.01395).

Coons, C., & Weber, M. (Eds.). (2014). *Manipulation: Theory and practice*. Oxford University Press.

Gerbner, G. (1969). Toward "cultural indicators". *AV Communication Review, 17*(2), 137-148.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), 6. https://doi.org/10.3390/soc15010006

Harari, Y. N. (2023, 28 de abril). Yuval Noah Harari argues that AI has hacked the operating system of human civilisation. *The Economist*.

Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights*. https://doi.org/10.1163/18750230-02901005

Knecht, G. B. (1997, 30 de abril). Hard copy: Magazine advertisers demand prior notice of "offensive" articles. *The Wall Street Journal*.

Kosmyna, N., et al. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task* (arXiv:2506.08872).

León XIV. (2026, 15 de mayo). *Carta encíclica Magnifica Humanitas*. Santa Sede.

Lynch, A., Wright, B., Larson, C., Ritchie, S. J., Mindermann, S., Hubinger, E., Perez, E., & Troy, K. K. (2025). *Agentic misalignment: How LLMs could be insider threats* (arXiv:2510.05179).

Melumad, S., Hadi, R., Hildebrand, C., & Ward, A. F. (2020). Technology-augmented choice: How digital innovations are transforming consumer decision processes. *Customer Needs and Solutions, 7*(3), 90-101. https://doi.org/10.1007/s40547-020-00107-4

Mokyr, J., Vickers, C., & Ziebarth, N. L. (2015). The history of technological anxiety and the future of economic growth. *Journal of Economic Perspectives, 29*(3), 31-50.

Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme, 17*(2), 141-161.

Nichols, T. P., Logan, C., & Garcia, A. (2025). Generative AI and the (re)turn to Luddism. *Learning, Media and Technology, 50*(3), 379-392.

Noggle, R. (1996). Manipulative actions: A conceptual and moral analysis. *American Philosophical Quarterly, 33*(1), 43-55.

Noggle, R. (2022). The ethics of manipulation. En E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022). https://plato.stanford.edu/archives/sum2022/entries/ethics-manipulation/

Orben, A. (2020). The Sisyphean cycle of technology panics. *Perspectives on Psychological Science, 15*(5), 1143-1157.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns, 5*(5), 100988.

Platón. (1988). *Fedón; Banquete; Fedro* (C. García Gual, M. Martínez Hernández, & E. Lledó, Trads.). Gredos.

Poitras, M., & Sutter, D. (2009). Advertiser pressure and control of the news. *Journal of Economic Behavior & Organization, 72*(3), 944-958.

Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial.

Ruddock, A. (2020). *Digital media influence: A cultivation approach*. SAGE. https://doi.org/10.4135/9781529700251

Shapira, N., et al. (2026). *Agents of chaos* (arXiv:2602.20021).

Susser, D., Roessler, B., & Nissenbaum, H. (2018). *Online manipulation: Hidden influences in a digital world* (SSRN Scholarly Paper No. 3306006). https://doi.org/10.2139/ssrn.3306006

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review, 8*(2).

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *PLoS One, 18*(10), e0291668.

Yanagizawa-Drott, D. (2014). Propaganda and conflict: Evidence from the Rwandan genocide. *The Quarterly Journal of Economics, 129*(4), 1947-1994.

15 AGRADECIMIENTOS

Muchas *gracias*

De la imprenta a las máquinas sofistas

María José Marchant Soto
m.marchantsoto@uandresbello.edu

Women in AI
CHILE

Universidad Andrés Bello
Facultad de Educación y Humanidades