Euskal Herriko Unibertsitatea
Universidad del País Vasco

GIZARTE ETA KOMUNIKAZIO ZIENTZIEN FAKULTATEA
FACULTAD DE CIENCIAS SOCIALES Y DE LA COMUNICACIÓN

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

El impacto de las Inteligencias Artificiales generativas en los procesos de deliberación pública

Autor: Jon Surja Herranz
Grado: Doble Grado en Ciencia Política y Gestión pública y Sociología
Curso 2025/2026

INDEX

INTRODUCCIÓN

La tendencia en estos últimos años relativa a la conversación pública ha visto cómo se poblaba de un nuevo interlocutor, que no son ni ciudadanos, ni políticos, ni activistas, ni periodistas, ni ningún actor que fuera objeto de conocimiento previo. Estos nuevos sujetos participantes son los sistemas automáticos de generación de lenguaje. Dichos sistemas intervienen en las conversaciones, responden preguntas, opinan sobre asuntos públicos, y en ocasiones, generan contenidos que son posteriormente consumidos y compartidos como si hubieran sido producidos por una mente humana. Herramientas como ChatGPT, Bard, o los integrados Meta AI y Grok en Whatsapp y X (antes Twitter), respectivamente, modifican silenciosamente el paisaje discursivo en el que se forma la opinión pública.

Tanto así, que este fenómeno ha despertado un interés creciente tanto en la academia como en el periodismo y la esfera política. Sin embargo, es notable como buena parte de la atención se ha centrado en cuestiones como la productividad, la eficiencia o los riesgos existenciales a largo plazo. Esta atención a las partes mencionadas le ha restado atención a la cuestión de qué ocurre cuando estos sistemas se integran de manera masiva en un nivel de deliberación pública (el de las redes sociales), mientras que el otro nivel (el de los parlamentos e instituciones) permanece relativamente ajeno a ellos. Si bien es una pregunta más sutil, es probable que sea más urgente también tener una respuesta. Entonces, ¿qué consecuencias tiene esta implantación asimétrica para la relación entre la política institucional y el discurso ciudadano?

La hipótesis que va a guiar la presente investigación es que la presencia creciente de inteligencia artificial generativa en el espacio público digital tiende a ensanchar el abismo que separa ambos niveles deliberativos. Por tanto, simplificando, la hipótesis es que esta presencia no es neutral. Además, esta disociación va más allá de que responda a cuestiones técnicas o a diferentes grados de adopción tecnológica, y es que responde a la diferencia que existe entre dos regímenes de producción de significado. Esta diferencia, que es más profunda, viene a señalar cómo mientras que la deliberación parlamentaria descansa sobre la capacidad humana de ofrecer explicaciones causales, de asumir responsabilidad por los enunciados y de anclar el lenguaje en la experiencia del mundo, los sistemas de IA generativa operan mediante correlaciones estadísticas entre formas lingüísticas. Esto es, no tienen acceso a la intencionalidad, al mundo o a la responsabilidad.

Para desarrollar esta hipótesis, el trabajo se estructura en tres grandes bloques. El primero, el marco teórico, expone los fundamentos conceptuales que permiten comprender qué clase de artefactos son estos sistemas y cuáles son sus límites. Se recorren para ello las aportaciones de la filosofía de la mente (Searle, Harnad), la lingüística computacional (Chomsky, Boleda) y la crítica tecnológica (Marcus, Bender, Weidinger, Valenzuela). El segundo bloque, el análisis, aplica estos fundamentos al objeto de estudio concreto, examinando cómo operan ambos niveles deliberativos, en qué consiste la asimetría entre ellos

y cuáles son sus consecuencias para la calidad del debate público y para la posición de la ciudadanía. Finalmente, las conclusiones recogen las principales implicaciones del análisis y apuntan algunas líneas para la reflexión normativa y la eventual regulación de estos sistemas.

Igualmente, el propósito de este trabajo no es, de ninguna manera, alimentar ninguna forma de tecnofobia ni de nostalgia por una edad de oro deliberativa que quizá habría que poner a debate en primer lugar. Se trata, más bien, de contribuir a una comprensión más precisa de los cambios que están teniendo lugar, de manera que tanto la ciudadanía como las instituciones puedan situarse ante ellos con una mayor claridad en sus funcionamientos. Porque, como se verá, el problema no es tanto que las máquinas se vuelvan demasiado inteligentes, sino que resultan lo bastante persuasivas como para que confiemos en ellas, y, presumiblemente, somos lo bastante ignorantes como para que esa confianza pueda ser peligrosa.

MARCO TEÓRICO

Para tratar de esclarecer el impacto de las Inteligencias Artificiales generativas en los procesos de deliberación pública habría que abordar primero qué clase de artefactos son estos sistemas y qué tipo de operaciones realizan cuando producen lenguaje. Durante las últimas décadas y especialmente en estos últimos años, es percibible cierta fascinación por los productos de IA, así como cierta perplejidad ante los mecanismos de estos modelos. Este marco teórico se construye, así, sobre la premisa de que la distinción fundamental no es cuantitativa, esto es, no es tanto cuánto procesan o qué tan rápido lo hacen, sino cualitativa. Es decir, que los grandes modelos de lenguaje (LLM) operan sobre correlaciones estadísticas entre formas lingüísticas, sin llegar a comprender el mundo como tal. Esta diferencia, que parece una cuestión técnica, tiene consecuencias políticas profundas e innegables cuando estos sistemas se integran en el espacio público y comienzan a modelar el discurso de la ciudadanía y la deliberación pública.

Respecto a la naturaleza de la comprensión, es necesario mencionar que la filosofía de la mente ha derivado una distinción a la hora de hablar de inteligencia artificial. Searle (1980) diferencia entre la intencionalidad intrínseca, propia de los seres biológicos dotados de cerebros con poderes causales específicos, y la intencionalidad derivada, que es aquella que los observadores proyectan sobre objetos que carecen de estados mentales. Por trasladarlo a un caso práctico, un libro contiene significado para quien lo lee, pero no para sí mismo. Del mismo modo, un ordenador ejecuta instrucciones formales sobre símbolos, pero no hay nadie al otro lado de esos símbolos que entienda qué significan. Entonces, podríamos decir que el argumento de la habitación china no ha perdido un ápice de vigencia, ya que un agente puede manipular símbolos chinos según reglas sintácticas, pasar el test de Turing y, sin embargo, no comprender ni una palabra de chino (Searle, 1980).

Harnad (1989, 1990, 1994) extrae de este argumento una consecuencia lógica y empírica. De la observación de que un sistema ejecuta un programa y no lo comprende, entonces ningún sistema que ejecute ese mismo programa comprende por el mero hecho de ejecutarlo. Por tanto, la comprensión resulta ser una propiedad del sustrato causal que lo ejecuta, y no una propiedad del programa. A partir de ahí, Harnad notifica el problema del anclaje de los símbolos, entendiendo este problema como que un sistema puramente simbólico es como un diccionario de un sólo idioma. Se refiere con esto a que los símbolos tienen que significar algo de manera intrínseca, porque si no, igual que el diccionario monolingüe, nos encontramos ante un bucle sin salida donde cada definición lleva a otra definición infinitamente. Entonces, los símbolos deberían estar anclados en representaciones no simbólicas, tales como las icónicas y categóricas, que conecten causalmente con los objetos que nombran (Harnad, 1990). El proceso humano del cerebro proporciona ese tipo de anclaje, gracias a su historia de interacción sensorial con el entorno. Pero los modelos de lenguaje, por diseño, carecen de esta raíz.

Otros autores, como Chomsky, Roberts y Watumull (2023) se han encargado de actualizar la crítica señalada, señalando que los modelos de lenguaje constituyen sistemas de correlación, no de explicación. Volviendo a la mente humana, esta no opera como motor estadístico, sino que lo hace de manera sorprendentemente eficiente, y de manera que con una cantidad de información reducida pueda construir explicaciones causales. A parte, también distingue lo posible de lo imposible, y ejerce una crítica creativa sobre sus propias conjeturas. Una niña aprende la gramática lingüística de su lengua nativa con una fracción de los datos que un modelo de lenguaje necesita para llegar a producir enunciados plausibles. Por la parte de la niña esto sucede porque ésta desarrolla un sistema de principios lógicos. Por otra parte, el modelo de lenguaje desarrollará una tabla de frecuencias. También hay que decir que los LLM aprenden con igual facilidad lenguas humanamente posibles e imposibles, ya que no hay dentro de su estructura algo que se corresponda con la distinción entre lo que es y lo que no es. No sabe, o por lo menos le cuesta, distinguir entre lo verdadero y lo falso, o entre lo permitido y lo prohibido.

Marcus (2022, 2023) ha insistido en la importancia de esta ausencia de comprensión, ya que ni es un defecto menor ni se trata de una fase transitoria en el desarrollo de la tecnología. De acuerdo con el autor, los sistemas actuales de deep learning no son capaces de superar el tamaño de los modelos, en tanto que más parámetros, más datos y un mayor coste computacional podrán mejorar la fluidez superficial, pero no la comprensión genuina. Es por eso que los problemas en referencia a confabulaciones de hechos inexistentes y a la incapacidad para mantener el rastro de las vinculaciones entre entidades y propiedades, que se le achacaban a GPT-2, siguen persistiendo en GPT-4. Así, Marcus traslada esto al ejemplo práctico de que estos sistemas se asemejan a un actor de improvisación que nunca se sale del personaje, y tampoco tiene experiencia real de la que hablar. Por tanto, estaríamos hablando de que la producción de estos sistemas es un collage estadístico de fragmentos textuales vistos durante el entrenamiento. También, posteriormente, estos fragmentos se recombinan mediante paráfrasis y sustitución de sinónimos.

Más allá de que este diagnóstico puede resultar filosóficamente curioso, no parece desacertado intuir que tiene consecuencias directas para cualquier análisis posterior que pretenda evaluar el papel de los grandes modelos de lenguaje en la deliberación pública. Si, como se ha mencionado antes, un sistema no distingue entre una afirmación verdadera y una falsa, o entre un argumento válido y una falacia, ¿cómo es posible que este mismo sistema pueda ser considerado un participante deliberativo en el mismo sentido en que lo es un ciudadano o representante político? Esto plantea interrogantes sobre su papel en la deliberación. Interrogantes que pueden salir con el uso creciente de la herramienta de X (anteriormente Twitter), Grok.

Una vez abordado el aspecto de qué tipo de máquinas son estos sistemas de inteligencias artificiales generativas, y cómo operan a la hora de producir lenguaje, toca comprender por qué estos sistemas producen lenguaje coherente aún sin comprender su significado. Para ello, es necesario examinar los fundamentos lingüísticos y de computación sobre los que se construyen. Por tanto, hay que analizar la semántica distribucional. Harris

(1954) formula originalmente la hipótesis distribucional. En esa hipótesis se sostiene que la similitud en el significado se corresponde con la similitud en la distribución lingüística. Más tarde, en 2020, Boleda explica que esta hipótesis, trasladada a su versión operativa, posibilita el desarrollo de modelos que representan las palabras como vectores en espacios geométricos de cientos de dimensiones. Así, la similitud semántica se mide por la cercanía de estos vectores. Dicho de otra manera, dos palabras son parecidas si aparecen en contextos parecidos.

Gracias a esta interpretación, se ha podido avanzar significativamente en tareas de recuperación de información, clasificación de documentos y traducción automática. Sin embargo, según la propia Boleda (2020), modelar el significado de una expresión resulta diferente a modelar el significado del hablante. Esto sucede porque la semántica distribucional no captura intenciones comunicativas. Puede capturar patrones de concurrencia, pero no intenciones de comunicación. Aterrizando esta idea, un vector podría reflejar que en un corpus dado "inmigración" aparece frecuentemente junto a "ilegal", pero no sería capaz de determinar si esa asociación responde a un hecho estadístico, a un sesgo del corpus o a una estrategia discursiva deliberada. En este punto conviene recordar que los datos no son entidades neutrales que flotan en el vacío. Como se ha señalado, los datos son constructos sociales, y constituyen constructos epistemológicos edificados sobre una determinada manera de entender el mundo. Por tanto, cualquier trabajo que sea el resultado de datos estará influido por esa manera de entender el mundo. Esto no es positivo ni negativo, es un hecho que debemos tener presente al hablar de datos (Innerarity, 2025; De la Iglesia, 2026).

Entonces, Baroni y Boleda (2019) analizan las diferentes arquitecturas de los modelos distribucionales y ofrecen una profundización en esta distinción. Los contextos de ventana corta son habituales en modelos entrenados con grandes cantidades de texto no filtrado. Estos capturan asociaciones temáticas amplias, pero son ciegos a las relaciones gramaticales finas. Así, los contextos de dependencias sintácticas, más sofisticados, pueden capturar relaciones de sujeto-objeto o de modificación adjetival, pero requieren corpus anotados y son computacionalmente más costosos. El caso de la mayoría de los LLM actuales, incluyendo aquellos que alimentan herramientas como Grok o ChatGPT, es que utilizan variantes de contextos de ventana. La consecuencia de esto es que esos modelos son especialmente sensibles a las palabras y patrones de alta frecuencia, siendo también notablemente insensibles a la estructura argumentativa compleja, las subordinaciones extensas y los matices modales. Estas cuestiones, imperceptibles para los modelos de lenguaje expuestos, son algunos elementos que caracterizan al discurso político deliberativo.

Después, los word embeddings, como señalan Baroni y Boleda (2019), transforman vectores one-hot ortogonales, en los que las palabras "motel" y "hotel" no guardan ninguna relación, en vectores densos donde la similitud geométrica codifica la similitud semántica. Este paso de la ortogonalidad a la continuidad tiene una contrapartida política, y es que un sistema que representa el significado como posición en un espacio continuo, no tiende naturalmente a tratar las diferencias políticas como diferencias de clase. En su lugar, las trata

como diferencias de grado. Es posible que posiciones que en el debate público son irreconciliables puedan ser acercadas por esta cuestión, de la misma manera que posiciones que comparten fundamentos comunes pueden ser alejadas. Se suele citar, como prueba de la potencia del modelo, la analogía vectorial rey - hombre + mujer = reina. Sin embargo, cabe destacar también que el mismo mecanismo puede generar, con los corpus adecuados, Putin - Rusia + Ucrania = invasor o inmigrante - trabajo + frontera = ilegal. Esto quiere decir, que el sistema no valida estas analogías, y genera cierta vulnerabilidad al extraer las analogías de los datos y estabilizarlas.

Marcus (2022) ha señalado la escasa fiabilidad de los LLM para entornos de alto riesgo, especialmente dada esta dependencia de las correlaciones estadísticas. Un sistema que no puede distinguir entre una afirmación verdadera y una confabulación plausible no debería ser utilizada en contextos como diagnósticos médicos, conducción autónoma, asesoramiento legal o asesoramiento psicológico. Esto, recalco, no responde a ninguna conspiración, sino a que algo que no distingue entre "bien" y "mal" estará habitualmente más cerca de ser un peligro que una herramienta. En el contexto del discurso político se puede trazar un símil. No está, por tanto, mejor informada una ciudadanía que recibe inputs generados por sistemas que no distinguen entre un hecho verificado y una invención estadística. Sin embargo sí que puede estar expuesta a una desinformación estructural que no requiere mala fe, sino un mero funcionamiento normal del sistema.

A partir de aquí, se pueden explorar los riesgos estructurales de los grandes modelos de lenguaje en el espacio público. Weidinger et al (2021) han elaborado una taxonomía completa hasta la fecha de los riesgos asociados al despliegue de LLM. El análisis de estos autores distingue seis áreas. La primera es la de la discriminación, la exclusión y la toxicidad. La segunda sería la de los peligros de la información. La tercera sería la desinformación. La cuarta los usos maliciosos. La quinta se refiere a los daños en la interacción humano-computadora, y la sexta se dirige a la automatización, acceso y daños ambientales. De alguna manera, cada una de estas áreas tienen implicaciones directas para la deliberación pública.

Los modelos de lenguaje no calibran, y reflejan el lenguaje de sus datos de entrenamiento. Si esos datos contienen estereotipos, discurso de odio o normas excluyentes, los modelos los aprenden y los amplifican. Weidinger et al. (2021) documentan que GPT-3 asocia "musulmán" con "terrorista" en un porcentaje significativo de casos, y que los nombres femeninos generan historias sobre familia y apariencia física con mayor frecuencia que los masculinos. Y esto, lejos de tratarse de un fallo del sistema, es su modo de funcionamiento normal. Constituye así una decisión política la elección del corpus de entrenamiento, ya que determinará qué voces serán representadas, qué asociaciones serán reforzadas y qué identidades serán invisibilizadas o patologizadas.

Hablando de deliberación pública, esta característica mencionada tiene consecuencias que van más allá de la mera reproducción de sesgos. Bender et al (2021) dan un paso más, y a que los modelos de lenguaje reflejan normas excluyentes, le añaden que las codifican en

forma de distribuciones de probabilidad que luego son utilizadas para generar texto. Así, una norma social que en un momento dado es mayoritaria, queda congelada en el modelo y es reproducida una y otra vez, aunque la sociedad haya avanzado ya a concepciones más inclusivas. Un ejemplo de esta secuencia puede ser la manera en la que se piensa la composición de la familia, y la presunción de que las familias están formadas por un padre, una madre e hijos biológicos. Este fenómeno ha sido denominado como el fenómeno de value lock-in por Gabriel y Ghazavi (2021). Y convierte a los LLM en fuerzas conservadoras por su arquitectura, que les inhabilita de distinguir lo que es estadísticamente frecuente y lo que es normativamente deseable.

Dado este marco, la desinformación se presenta como una consecuencia inevitable del desajuste entre el objetivo del entrenamiento y la naturaleza de la verdad (Weidinger et al. 2021). Por tanto, estamos hablando de algo que va más lejos de un problema de bulos, estamos hablando de que se produce un desajuste al maximizar la verosimilitud de las secuencias y la naturaleza de la verdad, que es una relación con el mundo y no una propiedad estadística de los enunciados. Los modelos de lenguaje tienen acceso a textos que hablan del mundo, pero no al mundo. De esta manera, puede aprender que los pájaros vuelan, pero también puede aprender que los pájaros no vuelan si esa negación aparece con suficiente frecuencia en su corpus. La frase afirmativa “Pedro Sánchez es presidente” y la afirmación “Pedro Sánchez no es presidente” pueden ser generadas con igual fluidez. El modelo carece de los mecanismos necesarios para contrastar enunciados con hechos, ya que la corrección fáctica no es una señal presente en los datos.

En este sentido, Bender y Koller (2020) sentencian que ningún sistema entrenado exclusivamente en forma lingüística puede adquirir comprensión del significado, porque el significado no está contenido en la forma. Para demostrar este límite teórico, proceden con el experimento mental del pulpo. Un pulpo que observa las conversaciones entre dos humanos en la superficie puede llegar a predecir con alta precisión qué dirá uno en respuesta al otro, pero si los humanos empiezan a hablar de construir una catapulta o de defenderse de un ataque de oso, el pulpo no es capaz de relacionar de qué están hablando. Esto se debe a que el pulpo jamás ha visto una catapulta o un oso, principalmente porque no tiene acceso al mundo al que las palabras se refieren. Los grandes modelos de lenguaje constituyen el pulpo del experimento.

Otros autores, como Valenzuela et al (2024) han reparado más en la capacidad de estos sistemas para constreñir la experiencia humana, reducir el espacio de posibilidades y transferir agencia a los usuarios de los algoritmos. Esto constituye un fenómeno más sutil que los riesgos catastróficos o existenciales de la IA. Su análisis identifica tres mecanismos que operan en este proceso. Además, lo hacen de manera simultánea y reforzandose mutuamente.

El primer mecanismo es el de la transferencia de agencia. ¿Cómo sucede esto? En tanto que el sistema de recomendación basa sus sugerencias en el comportamiento pasado del usuario, se crea un bucle de retroalimentación que refuerza las elecciones previas y dificulta la exploración de alternativas. Esto lleva a la pérdida de la posibilidad de encontrar algo

relevante sin haberlo buscado (serendipia), pero también erosiona la capacidad misma de cambiar de opinión. Una persona a la que el algoritmo alimenta constantemente con contenido de una determinada orientación política potencialmente verá reforzadas sus opiniones, pero además perderá práctica en contrastarlas con otras. Esto impactará en el hábito de entender los argumentos ajenos y en sostener la contradicción sin experimentarla como una amenaza.

Marcus (2022) ha observado que este fenómeno se extiende a toda interacción con IA generativa. La persona usuaria de ChatGPT que le delega la redacción de un argumento político está, pues, renunciando a ejercitar sus propias habilidades retóricas y analíticas. El sistema de lenguaje ofrece una alternativa rápida, fluida, inmediata y aparentemente competente. Esto puede derivar en pensar que el esfuerzo por pensar resulta superfluo. Alternativamente, esta situación puede plasmarse como situación de "deuda cognitiva". Este término se refiere a una delegación excesiva de nuestras habilidades cognitivas en el empleo de la inteligencia artificial, que en el largo plazo da lugar a una atrofia competencial y pérdida de dichas habilidades. Adicionalmente, Valenzuela et al. (2024) denominan este fenómeno como la descualificación, y no es tanto un efecto secundario indeseado, sino más bien la consecuencia lógica de un sistema diseñado para maximizar la eficiencia en la producción de texto.

El segundo mecanismo es el reduccionismo paramétrico. Como se ha mencionado antes, los algoritmos necesitan traducir la complejidad de la experiencia humana a un conjunto finito de variables computacionalmente tratables. No entienden los algoritmos acerca de identidades cambiantes, preferencias contradictorias o valores en tensión. Y esta traducción de complejidades deriva siempre a una reducción. La persona usuaria queda clasificada en edad, ubicación, historial de clics, interacciones previas… Posteriormente, sobre la base de ese perfil, el algoritmo decide qué contenido es relevante, qué información merece ser vista y qué argumentos merecen ser silenciados. Esta dinámica supone una objetivación con consecuencias en forma de oportunidades desiguales, discriminación automatizada y exclusión de aquellos cuyas características no encajan en las categorías previstas.

Sobre esto, Weidinger et al. (2021) documentan cómo esta objetivación afecta desproporcionadamente a grupos ya marginados. Los hablantes de inglés vernáculo afroamericano ven sus enunciados clasificados como más tóxicos que los de hablantes de inglés estándar. Esta condición no se da porque su lenguaje sea más violento, sino porque los corpus de entrenamiento y los clasificadores de toxicidad han sido desarrollados mayoritariamente por y para hablantes del estándar. En otro ejemplo práctico, las personas no binarias se enfrentan a sistemas de resolución de correferencia que fuerzan a elegir entre él o ella. De la misma manera, los hablantes de lenguas con pocos recursos digitales no les constan a estos modelos, y omiten su existencia. Generan texto fluido en inglés o en mandarín si se lo piden, pero raramente proporciona una respuesta coherente si se le pregunta algo elaborado en euskera, o en otros idiomas minoritarios.

El tercer y último mecanismo no puede ser otro que la expresión regulada. Los usuarios aprenden a modular su lenguaje para ser comprendidos por el sistema. Para ello, tienden a simplificar frases, eliminar ambigüedades y evitar ironías y metáforas complejas. En muchos casos estos comportamientos responden a una adaptación gradual a las limitaciones del interlocutor, más que a una elección consciente. Hildebrand, Hoffman y Novak (2023) han demostrado que las interacciones con asistentes de voz se caracterizan por un uso significativamente menor de pronombres personales, frases más cortas y una ausencia casi total de fórmulas de cortesía. Los usuarios hablan con Alexa de la misma manera en la que podrían estar dando órdenes a una máquina. No hablan como hablan con otro humano, porque la modalidad que la máquina puede procesar es la primera interacción, no la segunda.

Al mismo tiempo, esta regulación de la expresión tiene consecuencias que trascienden la interacción inmediata. Valenzuela et al. (2024) advierten que los usuarios pueden llegar a sentir que no pueden expresar su auténtica identidad a través de estos sistemas, o que el sistema amplifica ciertos aspectos de su personalidad mientras suprime otros. Una mujer que sigue cuentas de moda y belleza en Instagram verá su feed poblado de contenido sobre apariencia física, por mucho que su identidad profesional sea de abogada o ingeniera, y esto constituya un rol central para su autoconcepción. El algoritmo no sabe que ella es abogada, pero sabe que ha hecho clic en vestidos. Esta reducción de la persona afecta, en primer lugar, de manera clara, a los contenidos que ve. Pero en segundo lugar también afecta a cómo se percibe a sí misma.

La revisión de la literatura aquí presentada permite establecer una base conceptual sólida para abordar el objeto de esta investigación. De los autores examinados se desprenden tres ideas fundamentales: primera, que los grandes modelos de lenguaje operan mediante correlaciones estadísticas y carecen de comprensión semántica e intencionalidad intrínseca (Searle, 1980; Harnad, 1990; Chomsky, Roberts y Watumull, 2023); segunda, que su funcionamiento, basado en la semántica distribucional, reproduce y amplifica los sesgos presentes en sus datos de entrenamiento (Boleda, 2020; Bender et al., 2021); y tercera, que su despliegue en el espacio público conlleva riesgos estructurales de discriminación, desinformación y constricción de la experiencia humana (Weidinger et al., 2021; Valenzuela et al., 2024).

ANÁLISIS

Sobre la base conceptual expuesta en el marco teórico, el análisis que sigue se propone examinar cómo estos rasgos de la IA generativa se manifiestan de manera diferenciada en los dos niveles de deliberación pública, y qué consecuencias tiene esa implantación asimétrica para la relación entre la política institucional y el discurso ciudadano. Si en el centro del análisis se postula la disociación entre dos niveles de deliberación pública ante la irrupción de la IA generativa, la pregunta que atraviesa esta investigación podría plantearse de manera tal que: ¿Qué ocurre cuando un tipo de tecnología, diseñada para operar mediante correlaciones estadísticas y carente de comprensión semántica, se implanta masivamente en uno de los dos niveles en los que se articula la deliberación pública contemporánea, mientras que el otro nivel permanece relativamente ajeno a ello? La hipótesis que se sostiene es que esta implantación asimétrica, en primer lugar, no es neutral, y en segundo lugar, tiende a ensanchar el abismo que separa ese discurso de la política oficial desarrollada en los parlamentos. Nos encontramos en un momento histórico en el que vemos como en plataformas como X (antes Twitter) la presencia creciente de IA generativa deriva en unas cantidades ingentes de usuarios respondiendo ante cualquier hecho con "@Grok es esto cierto?". En hechos más flagrantes y denunciables, recientemente no pocos usuarios de manera pública pedían a Grok que le quitase ropa a personas (mayormente mujeres), y la IA respondía cumplimentando las solicitudes. Este fenómeno no es transitorio ni puede resolverse con mejoras técnicas, y la disociación entre niveles es resultado de una incompatibilidad ontológica entre dos regímenes de producción de significado.

De hecho, la cuestión de Grok y los desnudos públicos es otro ejemplo sobre cómo la IA generativa no calibra y no es capaz de distinguir entre "bien" y "mal" al ser solamente una serie de vectores. Distinguirá el análisis 4 secciones (1-4) con un subapartado en la sección 3, sobre los que se intentará trazar la situación en ambos niveles de deliberación pública, así como las consecuencias ontológicas de la cuestión.

1. LA DELIBERACIÓN PARLAMENTARIA COMO RÉGIMEN DE EXPLICACIÓN CAUSAL Y RESPONSABILIDAD INSTITUCIONAL

Comencemos por el primer nivel. Los parlamentos son instituciones en las que se procesan los conflictos sociales y el devenir del Estado mediante la palabra. En ellos, los representantes electos, siguiendo las formalidades de debate en cuanto a tiempos de intervención, turnos de palabra y turnos de réplica, deliberan sobre leyes, presupuestos y políticas públicas. Sin embargo, nos encontramos en un punto en el que hay que hacer un inciso sobre que más allá de esas reglas formales, la deliberación parlamentaria descansa sobre el presupuesto de que quienes participan en ella son seres humanos. Seres humanos capaces de ofrecer explicaciones causales, de justificar sus posiciones apelando a principios, de asumir responsabilidad por lo que dicen y de rendir cuentas ante sus electores.

Chomsky, Roberts y Watumull (2023) han insistido en que la diferencia distintiva entre un pensamiento humano y los LLM radica en la capacidad de explicar, y no meramente de correlacionar. Explicar, entendido no como predecir qué secuencia de palabras tiene más probabilidades de seguir a otra, sino como ofrecer razones, señalar causas, distinguir entre lo que es el caso y lo que podría serlo, o entre lo que sucedió y lo que habría sucedido si las condiciones hubieran sido diferentes. En esencia, es lo que haría un diputado que defiende una ley, ya que éste no se limita a ensamblar enunciados plausibles extraídos de su memoria. Este diputado construye un argumento que conecta principios generales con situaciones particulares, que anticipa consecuencias y que responde a objeciones. El argumento que construya el diputado será débil, interesado o incluso falaz, pero sigue siendo un argumento. Detrás de él hay alguien que lo sostiene y que responde por él.

A esta propiedad de los estados mentales humanos Searle (1980) la denominó como intencionalidad intrínseca. Esto es, que los pensamientos, las creencias y los deseos son acerca de algo, y lo son de manera directa, no porque alguien los interprete como tales. Si una diputada ofrece una rebaja de impuestos en una intervención está realizando un acto de habla que le obliga, que genera expectativas, y que dado el caso le podrá ser reclamada la rebaja. La obligación no deriva de que la persona oyente oiga una secuencia de sonidos y la interprete como una promesa, deriva, pues, de la intención del hablante y de las convenciones institucionales que le habilitan.

Este análisis se vio reforzado por Harnad (1990) cuando reportó que la intencionalidad requiere de un anclaje. Y es que los símbolos que maneja la mente humana, y las palabras que pronuncia, no significan nada por sí mismas. Estos y están adquieren un significado al estar conectados causalmente con el mundo a través de la propia experiencia sensoriomotora humana. Así el caso, otra diputada que hablase de pobreza, ha visto pobreza, o ha hablado con personas en situación de pobreza, o ha leído informes sobre pobreza. En definitiva, no son palabras que floten en un vacío semántico, porque están ancladas a alguna historia interactiva. Esta cuestión y este anclaje es lo que permite que esos enunciados sean vehículos de significado compartido y no un collage de fragmentos textuales.

Entonces, la deliberación parlamentaria va más allá de un intercambio de mensajes, y es un proceso en el que agentes intencionales, con vivencias en el mundo, anclados en el mundo y responsables de sus actos, producen explicaciones que aspiran a ser evaluadas por otros agentes intencionales. El proceso en el que se producen estas interacciones es posible que sea lento, o también es posible que haya formalismo y resistencia al cambio. ¿Considero estos rasgos como defectos? No. Los veo como condiciones de su funcionamiento, y no como defectos a corregir mediante la automatización. Una deliberación sin tiempo para la reflexión, que omita las garantías procesales a las minorías y que no permitiera cambiar de opinión con la ayuda de nuevos argumentos, difícilmente podría llamarse propiamente “deliberación”.

2. EL DISCURSO POLÍTICO EN REDES SOCIALES COMO SIMULACIÓN ESTADÍSTICA

Pasando al segundo nivel, el panorama cambia de manera radical. En las redes sociales predomina el discurso político que circula a velocidad vertiginosa. Además, lógicamente, no tiene las mediciones institucionales de las que el parlamento hace gala. En tanto que esas formalidades de turnos no existen, tampoco hay un registro oficial de lo dicho sobre lo que se puedan exigir responsabilidades, no de manera habitual, por lo menos. Hace aproximadamente un lustro empiezan a surgir herramientas, y resultan en que una parte creciente de ese discurso ahora mismo no es producido por humanos, sino por grandes modelos de lenguaje, a veces propiamente integrados en las plataformas mismas. Este es, por ejemplo, el caso de la herramienta Grok, que es el asistente de X (antes Twitter). Estas herramientas, como se ha mencionado antes, valen para generar texto político e incluso para participar en actos que implícitamente son políticos también. El segundo de los casos responde a que hubo cerca de un mes en el que Grok desnudaba a gente en la plataforma a petición de usuarios. En un formato más habitual, Grok es interpelado por los usuarios para verificar hechos. Sin embargo, queda de manifiesto en cada respuesta que es altamente maleable su “opinión”.

Al mismo tiempo que se habla del caso de Grok, se pueden asociar patrones similares con la IA integrada de Whatsapp, o el uso de la IA en LinkedIn. En cualquier caso, lo que parece que queda claro es que estos agentes no tienen capacidad de distinguir y propiamente deliberar.

¿Qué clase de agentes son estos sistemas? La respuesta, sostenida por un amplio consenso interdisciplinar, es que no son agentes en absoluto, al menos no en el sentido relevante para la deliberación. Tal vez puedan llegar a simular una deliberación, como estamos viendo de manera creciente con Grok. Pero no podrá llegar a deliberar. La cuestión, entonces, no es si los parlamentos deberían adoptar estas herramientas, sino qué ocurre cuando estas herramientas se despliegan masivamente en el espacio público y los ciudadanos las tratan como si fueran interlocutores válidos.

Weidinger et al (2021), en su taxonomía de riesgos, señalan que los LLM operan mediante la asignación de probabilidades a secuencias de tokens. Esto quiere decir que el objetivo de los modelos de lenguaje no es producir enunciados verdaderos, sino enunciados probables dada la secuencia anterior. Para ello, extraen patrones de corpus masivos que contienen todo tipo de textos. Estos textos pueden contener verdades, mentiras, argumentos, falacias, poesía, propaganda… Realmente, el modelo no distingue entre unas y otras al no tener acceso al criterio que permitiría hacerlo. No hay una base racional sobre la que sentar esos criterios.

Bender y Koller (2020) han demostrado que esta incapacidad, lejos de ser contingente, es incluso necesaria. El significado, argumentan, no es una propiedad de las formas lingüísticas, sino de la relación entre esas formas y el mundo. Un sistema entrenado

exclusivamente en formas no puede adquirir significado, porque la señal de entrenamiento carece de información sobre el mundo. Esto es independiente de lo grande que sea el corpus o lo grande que sea su arquitectura. Demuestran estos autores con su experimento mental del pulpo que el sujeto observador que solo ve intercambios lingüísticos entre dos humanos en la superficie puede potencialmente predecir con alta precisión sus respuestas, pero si los humanos empiezan a hablar de algo que el pulpo no ha visto nunca, como puede ser un partido de fútbol, la persona observante/el pulpo no es capaz de construir una idea de sobre qué están hablando.

La metáfora del loro estocástico, traída a este ensayo de la mano de Bender et al (2021), es una metáfora válida para explicar esta situación. Un loro puede repetir frases, incluso puede hacerlo en contextos que parecen apropiados, pero no sabe lo que dice. Del mismo modo, un modelo de lenguaje puede generar un discurso sobre la reforma fiscal que sea gramaticalmente impecable y estilísticamente convincente, pero no sabe qué es un impuesto, qué es la equidad fiscal o por qué alguien podría oponerse a una reforma. Sabe que en sus datos de entrenamiento las palabras impuesto, reforma y equidad aparecen juntas con cierta frecuencia, y explota esa información para producir secuencias plausibles. Pero el saber de un loro no es un saber.

Marcus (2022, 2023) ha documentado repetidamente las consecuencias de esta falta de comprensión. Los modelos confabulan hechos inexistentes con total naturalidad, a razón de que nada dentro de su arquitectura les advierte de que están inventando. Pueden recomendar el suicidio a pacientes ficticios, o también pueden inventar referencias bibliográficas falsas, y no lo hacen porque sean maliciosos, sino porque es su función. Como se ha dejado ver anteriormente, no están diseñados para maximizar la veracidad, en cualquier caso estarán para maximizar la verosimilitud. Esta verosimilitud en un corpus que contiene todo tipo de afirmaciones es compatible tanto con la verdad como con la mentira.

Marcus añade a este argumento una comparación con el caso de un actor de improvisación que puede sostener una conversación durante horas sin salirse del papel, pero al preguntarle por su experiencia real no tiene ninguna. Este actor que nunca rompe el personaje se asemeja a los modelos del lenguaje, ya que son actores con carácter perpetuo. No han votado nunca ni lo van a hacer, pero hablan de política, de leyes, de procesos de votación… Tampoco han sido juzgados nunca y son igualmente capaces de asumir que sí han sido juzgados y contar su “experiencia”. Este discurso es una simulación sin original.

3.1 LA DISOCIACIÓN ONTOLÓGICA Y SU ASIMETRÍA; LA NATURALEZA DE LA ASIMETRÍA

Después de estas exploraciones, queda de manifiesto la articulación de tres cuerpos teóricos. Se ha presentado la deliberación parlamentaria, se ha presentado el discurso en redes como simulación, y previamente se ha hecho un recorrido por la crítica a la inteligencia artificial como comprensión, seguido por el análisis de la semántica distribucional y acabando por explorar la taxonomía de riesgos sociotécnicos. El objetivo de este recorrido era

construir una hipótesis sobre la disociación entre los dos niveles de deliberación pública que constituyen el objeto central de esta investigación.

Así, el primer nivel, el de la deliberación parlamentaria oficial, opera bajo constricciones institucionales y cognitivas. Constricciones institucionales en tanto que está sujeto a procedimientos, plazos, responsabilidades y códigos de conducta que no son negociables, y constricciones cognitivas en tanto que los actores y actrices que participan en él son seres humanos dotados de intencionalidad intrínseca. Esto último quiere decir que son seres capaces de anclar sus símbolos en una historia de interacción con el mundo, son responsables de sus enunciados ante las masas y electores, y son responsables ante la historia. Si bien este nivel puede llegar a ser lento, ineficiente, capturado por intereses particulares, o distorsionado por la mala fe, son precisamente su lentitud y su ineficiencia lo que manifiestan una serie de garantías. Y es que la deliberación genuina normalmente requiere tiempo para contrastar argumentos, o para sopesar evidencias, o para incluso cambiar de opinión a la luz de nuevas razones.

Moviéndonos al segundo nivel, el de la deliberación pública en redes sociales, se ha intentado establecer el marco de que esta deliberación opera bajo un régimen completamente diferente. Los enunciados que lo pueblan son producidos, en proporción creciente, por sistemas que no tienen intencionalidad intrínseca, que no pueden anclar sus símbolos en experiencia alguna, que no son responsables de lo que dicen porque no hay nadie al otro lado de esos enunciados. Son sistemas que su razón de ser son, precisamente, la fluidez, la velocidad y la capacidad de producir texto a escala industrial. Están diseñados para maximizar el engagement, no la veracidad. Así como están igualmente diseñados para mantener al usuario en la plataforma, no para informarlo. En un sentido general, se prima el generar reacciones emocionales por encima de someter argumentos a escrutinio.

Nos encontramos entonces ante una brecha y una disociación entre ambos niveles. Esta brecha, además, no puede cerrarse mediante mejoras técnicas incrementales. No es posible resolverla haciendo que los modelos de lenguaje distingan mejor entre hechos y ficciones, mayormente porque esa distinción no es accesible desde su arquitectura. La opción de entrenar los modelos de lenguaje con corpus más diversos y cuidadosamente curados tampoco resolvería esta cuestión de fondo, ya que esta diversidad en el corpus seguiría sin compensar la falta de experiencia del mundo. De la misma manera que añadir capas de supervisión humana derivaría nuevamente en no compensar la falta de experiencia en el mundo, no compensar la ausencia de comprensión en el núcleo del sistema, y seguirían siendo, en el fondo, un conjunto de colecciones de formas lingüísticas.

Por tanto, esta disociación se trata de una disociación ontológica. Mientras que un nivel opera en el registro de la explicación causal y la responsabilidad política, el otro nivel opera en el registro de la correlación estadística, la producción masiva y la opacidad algorítmica. Merece la pena destacar también la justificación normativa del primer nivel mencionado. En cualquier caso, la ciudadanía recibe inputs del segundo nivel simulando ser outputs del primero, lo que consigue dejarla atrapada entre estos niveles. El discurso

generado por la IA no es más que un parecido estadístico, pero se hace pasar por algo político, informado y deliberativo. Esta simulación, a pesar de no ser una identidad sustancial, compite de manera efectiva con el original en condiciones de asimetría. Esto es, produce en mayor cantidad, con mayor rapidez, y aún produce adaptado a las preferencias reveladas del usuario. Pero lo que no puede producir es verdad, compromiso o responsabilidad, que sí produce el nivel deliberativo clásico.

De acuerdo con Chomsky, Roberts y Watumull (2023), estos sistemas son constitucionalmente incapaces de equilibrar creatividad y constricción. Pueden sobre-generar produciendo a partes iguales frases de carácter veraz y de carácter falso, como también pueden infra-generar exhibiendo un no-compromiso con cualquier decisión y una indiferencia total ante las consecuencias. ¿A qué lleva esta discusión? A que las imperfecciones de la deliberación clásica parlamentaria son el esfuerzo colectivo de equilibrar creatividad y constricción, que es, precisamente, lo que no llegan a saber medir los modelos de lenguaje. Los actos de crear nuevas leyes, introducir nuevas políticas y sugerir nuevas interpretaciones deberían estar siempre dentro de los límites constitucionales y del respeto a los derechos fundamentales. Sobre todo, cabe destacar que tienen que estar dentro de la responsabilidad ante los ciudadanos.

Y cabe destacar esta última frase en tanto que la IA generativa no puede participar en ese esfuerzo porque no tiene acceso a la noción misma de límite. No sabe qué es una Constitución, qué es un derecho o qué es una promesa electoral. Puede usar esas palabras, puede combinarlas en enunciados gramaticalmente correctos, puede incluso generar textos que parezcan defender una posición constitucionalista o criticar una promesa incumplida. Pero no sabe de qué está hablando. Y lo que parece aún peor que esta situación, es que tampoco sabe que no lo sabe.

A pesar de lo redactado en este marco teórico y análisis, no es en absoluto la intención la de conducir a una conclusión derrotista. Lo que se ha intentado poner de manifiesto, es que la literatura existente a día de hoy, conduce a pensar que la disociación entre los dos niveles es de naturaleza ontológica. Entonces, al no ser meramente técnica, las soluciones no deberían buscarse exclusivamente en el perfeccionamiento de los modelos, sino también en la reconfiguración del espacio público y en el establecimiento de límites normativos a qué tipo de agentes pueden participar en la formación de la opinión pública. Tal vez haya que exigir una transparencia sobre el origen de los enunciados, una educación crítica de la ciudadanía, y la creación de espacios deliberativos protegidos de la colonización algorítmica. Pero, en última instancia, la regulación democrática de estas tecnologías son vías que este marco teórico ha intentado abrir, y que en cualquier caso deberán investigar futuras investigaciones.

3.2 LA DISOCIACIÓN ONTOLÓGICA Y SU ASIMETRÍA; CONSECUENCIAS DE LA ASIMETRÍA

Llegamos así al núcleo de la disociación. Mientras que la IA generativa se está implantando masivamente en el segundo nivel deliberativo, su presencia en el primero sigue

siendo marginal. Esta asimetría entre su presencia en redes sociales y la deliberación parlamentaria en su modelo clásico responde a las diferentes exigencias de cada nivel.

Los parlamentos exigen fiabilidad, trazabilidad y responsabilidad. Por tanto, un sistema de lenguaje que no es capaz de distinguir entre la verdad y la mentira y al que no se le puede exigir que rinda cuentas de lo que dice, no tiene, en principio, cabida en ellos. Puede llegar a usarse como herramienta auxiliar, que no consta de momento que sea el caso, pero no debería ser un participante. Cuando se menciona "herramienta auxiliar" lo que se quiere decir es herramienta de resumen de documentos o redacción de borradores, en caso dado. Esta barrera no es una cuestión técnica, pero sí es institucional. Las normas que regulan la deliberación parlamentaria presuponen agentes intencionales y responsables. Un modelo de lenguaje no cumple esos requisitos, y ninguna mejora en su fluidez puede hacer que los cumpla.

Las redes sociales, en cambio, no exigen nada de eso. Su lógica es la del engagement, no la de la veracidad. Lo que importa es mantener al usuario en la plataforma, generar interacciones y recopilar datos. Un sistema que produce texto fluido a gran escala, que puede personalizar los mensajes según el perfil del usuario, que genera reacciones emocionales y alimenta la polarización, puede llegar a ser extraordinariamente valioso para ese propósito visto desde la órbita de un creador de contenido, particularmente. El hecho de que no entienda lo que dice es irrelevante, o incluso podría llegar a ser una ventaja, entendiendo que le está permitido decir cualquier cosa sin las inhibiciones que le impondría la responsabilidad.

Valenzuela et al (2024) aportan al análisis su profundización sobre cómo esta lógica afecta a la experiencia humana. Es relevante al análisis su concepto de transferencia. Este concepto se entiende como la delegación de un ciudadano en el algoritmo la selección de su dieta informativa, ya que no deja de estar renunciando a ejercer su propia capacidad de exploración. El algoritmo terminará por mostrar más contenido similar, reforzar sesgos y crear una cámara eco, haciendo desaparecer la posibilidad de serendipia. Así, el ciudadano se acaba encontrando en un círculo que se cierra sobre sí mismo al ver cómo el algoritmo elige sus opciones por él o ella.

La mencionada pérdida de serendipia puede tener sus consecuencias políticas directas, y es que cuando un ciudadano solo recibe información que confirma sus opiniones previas, su capacidad para comprender posiciones ajenas, o para matizar sus propias convicciones e incluso cambiar de opinión, se ve afectada y potencialmente se atrofia. No es tanto que se vuelva más radical, pero pierde la práctica del contraste y de la duda. El repertorio cognitivo se ve, pues, empobrecido, y se nota cuando se asoma al debate público. La deliberación requiere de la capacidad de ponerse en el lugar del otro, y de anticipar contraargumentos y sopesarlos, pero cuando las capacidades de contraste se ven mermadas, es más complicado llegar a esas posiciones citadas.

El reduccionismo paramétrico, otro de los mecanismos identificados por Valenzuela et al. (2024), agrava este problema. Por reduccionismo paramétrico se entiende el

funcionamiento de los algoritmos al reducir la complejidad de la identidad humana a un conjunto finito de variables. Sobre esa base de variables, que son la edad, el género, la ubicación, el historial de clics y las interacciones previas, se le asigna a cada usuario una posición en el espacio ideológico. En función de esa posición, a cada usuario se le sirve contenido adaptado. Pero esta reducción paramétrica tiende a ser una simplificación que puede resultar violenta, ya que el algoritmo no capta contradicciones, y un usuario que sea conservador en economía y progresista en costumbres, o puede estar a favor de la inmigración pero en contra de las cuotas. Por descontado, el usuario puede cambiar de opinión con el tiempo, pero el algoritmo aplana estas contradicciones y ofrece una versión simplificada del ciudadano. El ciudadano, por su parte, puede terminar sintiéndose identificado con ella al recibirla.

Weidinger et al. (2021) documentan cómo esta lógica afecta desproporcionadamente a grupos marginados. Los hablantes de dialectos no estándar ven sus enunciados clasificados como más tóxicos. Las personas no binarias se enfrentan a sistemas que las fuerzan a elegir entre dos géneros. Los hablantes de lenguas minoritarias sencillamente no existen para unos modelos entrenados mayoritariamente en inglés. Por lo tanto, estamos hablando también de que esta disociación ocurre tanto a niveles deliberativos como a nivel de grupos sociales. Unos grupos ven sus voces amplificadas por la IA, y otros grupos ven de alguna manera silenciadas las suyas por la misma estructura.

4. LA CONTAMINACIÓN DEL DISCURSO POPULAR Y EL ENSANCHAMIENTO DEL ABISMO

Después, cabe añadir también un efecto agregado sobre el discurso público. Se ha tratado de explicar cómo la presencia masiva de IA generativa en el segundo nivel deliberativo limita y constriñe la experiencia de los usuarios individuales. Pero también ocurre un efecto agregado sobre el discurso público en su conjunto, y ese efecto puede describirse como contaminación semántica.

Los LLM generan texto a una escala y velocidad inigualables para la mente humana. Pueden producir miles de comentarios, decenas de miles de publicaciones y una cantidad de texto notable en lo que una mente humana activista redacta un argumento. Esa misma capacidad de saturación altera la ecología del discurso. Los mensajes generados por IA compiten con los humanos por la atención de los usuarios, y poseen, de manera objetiva, unas ventajas referidas a velocidad y preferencias estadísticas que no son comparables realmente a la capacidad humana. Esto funciona, de manera añadida, particularmente bien en un mundo donde la lógica capitalista impulsa a pensar el tiempo como un valor en sí mismo (Mendiola, 2026).

Sin embargo, esta fluidez no les hace realmente más verdaderos. Como han señalado Weidinger et al. (2021), los modelos de lenguaje no tienen mecanismos para distinguir la información veraz de la falsa. Generan ambas con la misma fluidez. Los modelos de lenguaje tienen la capacidad de generar bulos sobre un político, teorías conspirativas sobre una

votación o noticias falsas, y pueden circular con una apariencia similar de legitimidad que un informe veraz y verificado. El usuario medio no tiene las herramientas para distinguir unos de otros, y la fuente habitualmente no es visible. Por tanto, el mensaje, independientemente de su origen, es lo que cuenta.

Volviendo momentáneamente a la velocidad y escala de producción de los LLM, habría que añadir un problema adicional, que es el que constituye el *Slop* ("bazof-IA", adaptado al castellano). Esto es, contenido digital y textual mediocre producido por la IA como resultado del procesamiento estadístico de la producción humana, sin ningún tipo de criterio ni verificación. Es carente de originalidad, calidad y atractivo, reproduciendo patrones convencionales, representando ideas banales, y en ocasiones absurdas, presentándose como una ola de producción de contenido mediocre no supervisado.

Chomsky, Roberts y Watumull (2023) advierten que estos sistemas son constitucionalmente incapaces de equilibrar creatividad y constricción. O bien lo aceptan todo, produciendo tanto verdades como falsedades, o bien no se comprometen con nada, mostrando indiferencia ante las consecuencias. Una diferencia plasmable con la deliberación democrática, por el contrario, es el esfuerzo por equilibrar ambas cosas. Las nuevas propuestas que surjan deben estar encuadradas dentro de los límites constitucionales y éticos. La IA no tiene acceso a la noción misma de límite, es por eso que no puede participar en ese esfuerzo.

El resultado de este desequilibrio resulta en un ensanchamiento entre la política oficial y el discurso popular. Los parlamentos siguen los procedimientos de deliberación clásica, lentos, procedimentales, sujetos a verificación y responsabilidad. Como se ha argumentado, estas son condiciones de posibilidad de la deliberación genuina, pero que ante el estándar de las redes aparecen como defectos. Las redes producen una cantidad enorme de mensajes rápidos, emocionales y que no están sujetos a verificación, ni mayormente a responsabilidad. La ciudadanía, estando expuesta a esta segunda corriente, puede desarrollar expectativas y percepciones que no se corresponden con la realidad de la primera. Así, situaciones de matización y ponderación, pueden traducirse en redes como situaciones "dubitativas" y "cobardes". El no ser compatible esta deliberación clásica con la velocidad de las redes, al mismo tiempo, parece otorgarle un carácter "lento". Pero esta simulación no compite realmente con el modelo original, ya que se encuentran en condiciones de asimetría.

Si bien el verificar los hechos reales con un modelo de lenguaje, o el pedirle a otro modelo que redacte los posts a publicar, conlleva a una producción de enunciados y argumentos más rápidos y más ajustados al estado emocional del momento, estos artefactos no pueden producir ni verdad, ni compromiso ni responsabilidad. La ciudadanía, al costarle distinguir entre ambos regímenes, termina por evaluar la política real con los criterios de la política simulada. Esto es lo que refleja, de manera clara, la enorme cantidad de usuarios de X (Twitter) tratando de verificar los hechos reales con Grok. De manera bastante previsible, el parlamento sale perdiendo en esa comparación. No porque sea peor que antes, sino porque el estándar con el que se le mide ha cambiado.

Marcus (2022) ha señalado que el peligro no proviene de una superinteligencia que nos esclavice, sino de sistemas lo bastante persuasivos como para que la gente confíe en ellos, y lo bastante ignorantes como para causar daño. La disociación que aquí se analiza es una manifestación de ese peligro, y es que a la IA, realmente no le hace falta la "I" para contaminar el discurso público. No le hace falta ser un modelo inteligente, le hace falta ser un modelo que produzca con fluidez, y le hace falta producir mensajes que parezcan políticos, si bien tampoco necesita comprender la política. Por tanto, podemos decir que en este momento, la apariencia llega a sustituir a la realidad, y la simulación se convierte en un régimen discursivo que la ciudadanía acepta.

CONCLUSIONES

Tras este recorrido analítico, las conclusiones quedan reservadas para preguntarse si esta disociación es reversible. La respuesta depende de lo que entendamos por reversibilidad. Si se trata de mejorar los modelos para que distingan mejor entre verdad y falsedad, la respuesta es negativa. Como han argumentado Bender y Koller (2020) y Harnad (1990), esa distinción no es accesible desde una arquitectura basada únicamente en formas lingüísticas. Dado que un modelo de lenguaje no va a poder tener experiencias sensoriomotoras, ni un anclaje en el mundo, ni una historia de interacciones causales. Y estas características son las condiciones de posibilidad del significado. Por tanto, sin ellas, el resultado del texto producido es un conjunto de palabras que contienen sintaxis, pero no significado.

Si, por otra parte, entendemos la reversibilidad como tratar de regular el uso de estos sistemas en el espacio público, se abre un espacio para discutir la posibilidad. Weidinger et al. (2021) mencionan varias vías para regular los modelos de lenguaje en la esfera pública. Se destacan en este análisis: La transparencia sobre el origen de los mensajes, las marcas de agua en el texto generado, las limitaciones a los usos maliciosos, y la participación de las comunidades afectadas en el diseño de los sistemas.

Estas propuestas de la literatura académica encuentran un correlato institucional en el desarrollo del marco regulatorio europeo. La Unión Europea ha desplegado en los últimos años un conjunto de normas que, de manera directa o indirecta, afectan al desarrollo y despliegue de los sistemas de inteligencia artificial. La siguiente tabla sintetiza las principales disposiciones y su relevancia para el fenómeno que aquí se analiza.

Regulation	Primary Focus/Scope	Relevance to AI
AI Act	Horizontal framework for AI using a risk-based tiering system.	Establishes the constitutional bedrock for all AI providers; includes strict extraterritorial reach (Art. 2) for systems impacting EU users.
Liability Rules	Modernization of product and AI-specific civil liability.	Codifies the principle of accountability, ensuring redress for damages caused by high-risk software and autonomous models.
Digital Services Act (DSA)	Governance of online intermediaries and platform systemic risks.	Provides the regulatory basis for algorithmic content moderation, recommendation systems, and transparency in platform governance.
Digital Markets Act (DMA)	Ex-ante obligations for systemic digital "gatekeepers."	Curbs the influence of Big Tech on digital infrastructure, preventing the monopolization of data sets essential for AI training.
Data Governance Act (DGA)	Mechanisms for secure data sharing and "data altruism."	Facilitates the availability of high-quality, trusted data sets for AI development while upholding public interest safeguards.
Data Act	Access and usage rights for data generated by connected products.	Rebalances control of IoT-generated data between manufacturers and users, breaking proprietary "data silos" for AI deployment.
EHDS Regulation	European Health Data Space for secondary data usage.	Enables the use of sensitive electronic health records for medical AI research under stringent privacy and security protocols.

(Fuente: Itziar Alkorta Idiákez, 2026)

Este entramado normativo constituye un reconocimiento, desde la esfera institucional, de que los riesgos asociados a la IA no son meramente técnicos sino que requieren respuestas políticas. La existencia de leyes como la AI Act o la Digital Services Act indica que la disociación entre niveles deliberativos no ha pasado inadvertida para los reguladores. Sin embargo, queda por ver si estas herramientas serán suficientes para proteger el espacio de la deliberación clásica sin la innovación, y si lograrán dotar a la ciudadanía de los instrumentos necesarios para distinguir entre la comunicación política auténtica y la simulación estadística.

Retomando las vías de transparencia, marcas de agua, limitaciones y participación de comunidades afectadas, propuestas por Weidinger et al (2021), otras investigaciones, como la de Valenzuela et al. (2024) añaden la necesidad de diseñar algoritmos que incorporen serendipia y que no reduzcan a los usuarios a perfiles unidimensionales.

¿Realmente está asegurado que incorporando estas medidas se resolverá la disociación ontológica entre los 2 niveles? Son medidas que pueden parecer necesarias, pero los parlamentos y la deliberación oficial no dejan de ser espacios de explicación causal y responsabilidad institucional, mientras que las redes carecen de estas dos características, y la producción y lectura de contenido político con IAs generativas son, en el fondo, un espacio de simulación estadística. Puede llegar a estrecharse el abismo entre ambos, pero parece difícil que desaparezca.

Entonces, lo que queda puede ser una pregunta más profunda, que no responde ni a cuestiones técnicas ni regulatorias. Políticamente, ¿qué tipo de deliberación queremos como ciudadanos? Si la respuesta es la deliberación clásica, con sus garantías y sus lentitudes, entonces habrá que proteger ese espacio de la colonización algorítmica. Si la respuesta es una deliberación híbrida, que integre la producción automatizada de discurso, entonces habrá que repensar desde la base qué significa deliberar, qué significa representar y qué significa rendir cuentas. No está claro que esas preguntas tengan una respuesta satisfactoria.

Mientras que estas preguntas quedan sujetas a respuestas, los parlamentos y las redes sociales siguen deliberando. Las redes sociales generan un discurso a velocidad aumentada, mientras la deliberación clásica sigue sus cauces institucionales. Así, la ciudadanía desarrolla una relación complicada en la que exige respuestas rápidas, que se ven desafiadas por la desconfianza de la lentitud institucional. Adicionalmente, puede haber indignación con la falta de transparencia, pero se sigue consumiendo contenido de origen opaco. En la misma línea, se reclama autenticidad pero se interactúa con simulaciones sin ninguna base real del mundo. En este espacio citado de contradicciones, es necesario esclarecer cómo la deliberación pública se ve afectada por estos procesos.

Conclusivamente, la inteligencia artificial generativa no puede, pues, presentarse como una solución a los problemas de la deliberación democrática. En cualquier caso, será un síntoma de un cambio y una transformación más profunda en la manera en que producimos, consumimos y legitimamos el discurso público. Ante este espacio transformativo, cabe

comprender esta transformación y tomar una decisión colectiva acerca de si queremos ser nosotros quienes empleemos la tecnología o si preferimos que ella nos emplee. Pero, tal y como se ha argumentado, el dejarse llevar por la corriente de una tecnología que no entiende lo que dice no parece la mejor solución.

BIBLIOGRAFÍA

- Alkorta Idiákez, I. (2026). AI REGULATION AND THE “BRUSSELS EFFECT”: GLOBAL IMPACT OF EU RULES ON AI TECHNOLOGY STANDARDS. Euskal Herriko Unibertsitatea (EHU/UPV).
- Baroni, M., y Boleda, G. (2019). Distributional semantics [Apuntes]. CS 388, Universidad de Texas en Austin.
- 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 (pp. 610-623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922
- Bender, E. M., y Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. En Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5185-5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463
- Boleda, G. (2020). Distributional semantics and linguistic theory. Annual Review of Linguistics, 6, 213-234. https://doi.org/10.1146/annurev-linguistics-011619-030303
- Chomsky, N., Roberts, I., y Watumull, J. (2023, 10 de marzo). AI unravelled: The false promise of ChatGPT. DT Next. https://www.dtnext.in/edit/ai-unravelled-the-false-promise-of-chatgpt-the-human-mind-is-not-like-chatgpt-and-its-ilk-a-lumbering-statistical-engine-for-pattern-matching-it-is-a-surprisingly-efficient-and-elegant-system-that-operates-with-small-amounts-of-information-it-seeks-not-to-infer-brute-correlations-among-data-points-but-to-create-explanations
- De la Iglesia, E. (2026). Technophilosophy: Rethinking Technology as a Human and Social Challenge. Mondragon Unibertsitatea.
- Gabriel, I., y Ghazavi, V. (2021). The challenge of value alignment: From fairer algorithms to AI safety. arXiv. https://doi.org/10.48550/arXiv.2101.06060
- Harnad, S. (1989). Minds, machines and Searle. Journal of Experimental and Theoretical Artificial Intelligence, 1(1), 5-25. https://doi.org/10.1080/09528138908953691
- Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3), 335-346. https://doi.org/10.1016/0167-2789(90)90087-6
- Harnad, S. (1994). Computation is just interpretable symbol manipulation: Cognition isn't. Minds and Machines, 4(4), 379-390. https://doi.org/10.1007/BF00974165
- Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162. https://doi.org/10.1080/00437956.1954.11659520
- Hildebrand, C., D. Hoffman, and T. Novak (2023). “Your Request Is My Command! How Initiation Modalities Shape Conversational AI Experiences,” Working Paper.
- Innerarity, D. (2025). Una teoría crítica de la inteligencia artificial. Galaxia Gutenberg
- Marcus, G. (2020, 22 de agosto). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. MIT Technology Review.

https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/
- Marcus, G. (2022, 11 de diciembre). Gary Marcus: "El gran problema de la inteligencia artificial actual es que no es fiable". El Confidencial. https://www.elconfidencial.com/tecnologia/2022-12-11/chatgpt-openai-gary-marcus-ia-ai-inteligencia-artificial_3537495/
- Marcus, G. (2022). Deep learning is hitting a wall. Nautilus. https://nautil.us/deep-learning-is-hitting-a-wall-238440/
- Marcus, G. (2022). How come GPT can seem so brilliant one moment and so breathtakingly dumb the next? https://garymarcus.substack.com/p/how-come-gpt-can-seem-so-brilliant
- Mendiola, I. (2026). *Aparición del capitalismo* [Diapositiva 7]. Dimensiones Sociales de la Economía. Euskal Herriko Unibertsitatea (EHU/UPV).
- Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417-424. https://doi.org/10.1017/S0140525X00005756
- Valenzuela, A., Puntoni, S., Hoffman, D., Castelo, N., De Freitas, J., Dietvorst, B., Hildebrand, C., Huh, Y. E., Meyer, R., Sweeney, M. E., Talaifar, S., Tomaino, G., y Wertenbroch, K. (2024). How artificial intelligence constrains the human experience. Journal of the Association for Consumer Research, 9(3), 241-255. https://doi.org/10.1086/730709
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv. https://arxiv.org/abs/2112.04359