EDICIONES COMPLUTENSE

**Teknokultura**. Revista de Cultura Digital y Movimientos Sociales
ISSNe: 1549-2230

KARPETA

# El reto de la moderación de contenidos en inteligencia artificial generativa: ChatGPT bajo el marco regulatorio de la Unión Europea

**Guillermo Prieto-Viertel**
Universitat Oberta de Catalunya (España)
**Raúl Tabarés Gutiérrez**
Fundación Tecnalia Research & Innovation (España)

https://dx.doi.org/10.5209/TEKN.97799 Recibido: 9 de septiembre de 2024 • Aceptado: 8 de marzo de 2025 • **REVISIONES EN ABIERTO**

ESP **Resumen.** Este artículo aborda la problemática de la moderación de contenidos en los sistemas de Inteligencia Artificial Generativa (IAG). A través del caso de ChatGPT, exploramos como los diferentes instrumentos regulatorios impulsados recientemente por la Unión Europea no abordan de manera explícita la moderación de los contenidos generados por la IAG. Además, se argumenta que las plataformas privadas siguen operando con un alto grado de discrecionalidad, lo que genera preocupaciones significativas sobre su capacidad para gestionar de manera justa y efectiva los riesgos inherentes a la IAG. El texto enfatiza la importancia de democratizar la gobernanza en torno a la moderación de contenidos para proteger a los usuarios sin comprometer las libertades fundamentales, así como la multiculturalidad del espacio en línea.
**Palabras clave:** censura; grandes modelos de lenguaje; inteligencia artificial; ley; reglamento de servicios digitales.

## ENG The challenge of content moderation in generative artificial intelligence: ChatGPT under the regulatory framework of the European Union

ENG **Abstract.** This article addresses the issue of content moderation in Generative Artificial Intelligence (GAI) systems. Through the case of ChatGPT, it explores how the different regulatory instruments recently promoted by the European Union do not explicitly address the moderation of content generated by GAI. Additionally, it is argued that private platforms continue to operate with a high degree of discretion, raising significant concerns about their ability to manage the inherent risks of GAI fairly and effectively. The text emphasizes the importance of democratizing governance around content moderation to protect users without compromising fundamental freedoms, as well as the multicultural nature of online space.
**Keywords:** artificial intelligence; censorship; digital services act; law; large language models.



**Cómo citar:** Prieto-Viertel, Guillermo y Tabarés Gutierrez, Raúl (2025). El reto de la moderación de contenidos en inteligencia artificial generativa: ChatGPT bajo el marco regulatorio de la Unión Europea, *Teknokultura. Revista de Cultura Digital y Movimientos Sociales 22*(2), 161-174. https://dx.doi.org/10.5209/TEKN.97799

## 1. Introducción

En la actualidad, la Inteligencia Artificial (IA) ha catalizado transformaciones digitales significativas en diversos sectores de actividad, que van desde el comercio y la comunicación hasta la creatividad y la cultura. El reciente desarrollo de la Inteligencia Artificial Generativa (IAG) permite la generación de textos, imágenes y otros medios en respuesta a las solicitudes de los usuarios, lo que crea tanto oportunidades extraordinarias como desafíos éticos y regulatorios relevantes. Ejemplos notables de estos sistemas incluyen ChatGPT de OpenAI, Gemini de Google y Claude de Anthropic, entre otros, que han emergido recientemente.

Aunque la moderación de contenidos en las plataformas digitales tradicionales ya supone un desafío complejo, la capacidad de la IAG para responder instantáneamente a prácticamente cualquier consulta introduce un reto único en cuanto a la moderación de contenidos. A diferencia de los contenidos creados por usuarios, en la IAG nos enfrentamos a contenidos generados automáticamente por algoritmos que pueden producir material inapropiado o problemático sin una clara intención humana detrás. Regular este tipo de sistemas no solo requiere una vigilancia continua, sino también un profundo replanteamiento de las estructuras regulatorias existentes, que no están preparadas para abordar las particularidades de la IAG (Wachter, 2024).

Recientemente, la atención se ha centrado en los peligros asociados con la generación de contenido falso por parte de la IA, tales como *deepfakes* o noticias falsas, así como la creación de contenido erróneo o incluso ilegal (Hacker et al., 2023, 2024; Wachter et al., 2024). Esta creciente preocupación ha impulsado un aumento en los esfuerzos por limitar y moderar las respuestas de la IAG para mitigar estos riesgos. Sin embargo, este exceso de moderación puede llegar a ser contraproducente, extralimitándose y resultando en una potencial censura relativa a diferentes contenidos controvertidos, poniendo en relieve la necesidad de encontrar un equilibrio entre la protección contra los riesgos de la IA y la preservación de la libertad de expresión y el acceso a la información.

A pesar de los avances regulatorios, aún persisten importantes lagunas en la normativa que rige la moderación en la IAG en la Unión Europea (UE) (Wachter, 2024). Por ello, es fundamental examinar el entramado legal y normativo que regula estas herramientas. Para ello, este artículo realiza un análisis exhaustivo de las regulaciones europeas existentes (Reglamentos de Servicios Digitales, Reglamento de Inteligencia Artificial y Directiva de Responsabilidad por Productos), con el objetivo de confrontar como abordan los retos que plantea la IAG. En particular, analizamos cómo ChatGPT se adhiere a estas regulaciones y cómo se moderan los contenidos generados por este sistema de IAG en diferentes situaciones, proporcionando además un análisis de la decisión de moderación y la revisión de los Términos de Uso. A partir de ejemplos de extralimitación en la moderación de contenidos en esta plataforma se argumenta cómo la falta de directrices claras genera inconsistencias en la aplicación de políticas de moderación, lo que puede derivar en restricciones injustificadas o en la eliminación arbitraria de ciertos discursos.

Para la selección de los ejemplos de extralimitación en la moderación de contenidos en ChatGPT, se llevó a cabo una búsqueda en foros y redes sociales, utilizando tanto los términos 'moderación' como 'censura'. Se aplicaron criterios de inclusión basados en la relevancia del caso, en relación con la problemática estudiada, y la disponibilidad de documentación visual o textual que permitiera un análisis detallado. La selección final de los ejemplos se realizó con el objetivo de incluir casos que reflejen diversas temáticas donde puedan evidenciarse lagunas regulatorias, así como instancias de extralimitación que correspondan a distintas secciones de los Términos de Uso de ChatGPT, permitiendo así un análisis más amplio de los desafíos en la moderación de contenidos

Paralelamente, la selección bibliográfica siguió un enfoque sistemático, priorizando literatura reciente debido a la rápida evolución de la regulación de inteligencia artificial y moderación de contenidos. Se consultó la base de datos Google Scholar, utilizando términos clave como 'moderación de contenidos', 'Inteligencia Artificial Generativa', 'reglamentos de servicios digitales' y 'reglamento de Inteligencia Artificial'. Además, se incluyeron informes normativos de la Unión Europea y se incorporaron referencias fundamentales sobre la moderación de contenidos en plataformas digitales para proporcionar un marco teórico.

Los resultados de este estudio sugieren que estas regulaciones europeas actuales no abordan de manera explícita la moderación de los contenidos generados por la IAG, lo que permite que las plataformas privadas operen con un alto grado de discrecionalidad (Gillespie, 2018; Tabarés Gutiérrez, 2024; York, 2022), Además, se identifican problemas como la falta de transparencia en los criterios de moderación, la tendencia a la extralimitación en la eliminación de contenidos y la dificultad de equilibrar seguridad y libertad de expresión. Estos desafíos evidencian la necesidad de desarrollar mecanismos de rendición de cuentas y supervisión externa, así como de impulsar investigaciones futuras para evaluar el impacto real de la moderación en la percepción de los usuarios y la protección de derechos fundamentales.

El artículo se estructura de la siguiente manera: a continuación, se provee de una revisión de la problemática de la moderación de contenidos en plataformas digitales, destacando los desafíos específicos que surgen en las plataformas de IAG. Posteriormente se hace una presentación y análisis de las diferentes regulaciones UE en este ámbito. En la cuarta sección se aborda el caso de estudio de ChatGPT, profundizando en las implicaciones prácticas de estas normativas para, posteriormente, proveer de una discusión y unas breves conclusiones.

## 2. Desafíos y retos de la moderación de contenidos

La moderación de contenidos enfrenta numerosos desafíos derivados de factores legales, éticos, técnicos y sociales. Las plataformas digitales deben equilibrar la aplicación de normas comunitarias con la protección de la libertad de expresión, ya que su responsabilidad de moldear y, a veces, restringir el contenido plantea importantes preguntas sobre la censura y el discurso público (Gillespie, 2018). Las diferencias culturales y legales complican aún más este equilibrio, porque lo que se considera dañino puede variar ampliamente entre regiones (York, 2022). La escala a la que operan plataformas digitales es vasta, lo que hace de la moderación de contenidos un desafío significativo. Como señaló Del Harvey, vicepresidenta de Confianza y Seguridad en Twitter, en su charla TED, una probabilidad de uno en un millón ocurre quinientas veces al día en Twitter. Esto significa que, incluso si el 99,999 % de los tuits no presenta ningún riesgo, el 0,001 % restante se

traduce en aproximadamente 150.000 tuits problemáticos por mes, lo que resalta el inmenso desafío que representa la escala del contenido (Harvey, 2014).

De manera similar, ChatGPT, con casi de 450 millones de usuarios mensuales y más de 2,6 mil millones de visitas en agosto de 2024 (SimilarWeb, 2024), participa en millones de conversaciones diariamente. Este volumen implica que, incluso las ocurrencias raras, como la generación de respuestas inapropiadas o incorrectas, pueden suceder con frecuencia. Por ejemplo, un abogado que utilizó ChatGPT para preparar un informe legal, que terminó incluyendo citas falsas, enfrenta una audiencia para discutir posibles sanciones debido a no verificar la veracidad de la información proporcionada (Prego, 2023). Otro caso es el del *chatbot* llamado Tessa que tuvo que ser desactivado después de proporcionar consejos perjudiciales a personas con trastornos alimentarios (Harper, 2023). Las personas pueden sentirse inclinadas a creer estas afirmaciones por varias razones: la autoridad que emana de las IAG, que son buenas informando con precisión en muchos contextos, y que las personas no comprenden cómo funcionan ni que pueden sufrir alucinaciones (Henderson et al., 2023). La gestión de tales interacciones requiere de algoritmos sofisticados y sistemas de monitoreo para garantizar que la IAG proporcione respuestas precisas, seguras y contextualmente apropiadas.

Existe una extensa literatura sobre la moderación de contenidos en plataformas digitales tradicionales, donde el enfoque principal ha sido gestionar el contenido subido por los usuarios (Gillespie, 2018; Gorwa et al., 2020; Tabarés Gutiérrez, 2024), como se detalla en la Tabla 1. Sin embargo, nos encontramos ante un nuevo paradigma en el contexto de la IAG. A diferencia de las plataformas tradicionales, donde los usuarios son los creadores de contenido, en las plataformas de IAG, los contenidos son generados automáticamente por algoritmos sin una intención humana detrás. Este cambio fundamental plantea nuevos desafíos y requiere enfoques innovadores para la moderación. Para empezar, las plataformas tradicionales dependen, en gran medida, de la intervención humana para garantizar que el contenido cumpla con las normas comunitarias y los estándares legales (Roberts, 2019). Sin embargo, la moderación en tiempo real por parte de humanos en el contenido generado por IA es impracticable debido a la naturaleza rápida y continua de las interacciones de la IA.

La IA también enfrenta desafíos en la interpretación del contexto, la ironía y la jerga. La necesidad de intervención humana resalta por las limitaciones de la IA, ya que es muy efectiva en la identificación de contenido explícito como pornografía y spam, pero no tan hábil en la identificación de discursos de odio o contenido dañino (Dias Oliva, 2020; Sorbán, 2021). Por ejemplo, las máquinas tienen dificultades para diferenciar entre frases similares con diferentes sentimientos, como 'tu puta madre' (negativo) y 'de puta madre' (positivo) (York, 2022). Este desafío se agrava con palabras extranjeras, donde el sentimiento debe considerarse en todos los contextos culturales y cuando las empresas a menudo priorizan idiomas principales, como el inglés, el español y el francés. Esta falta de apoyo afecta a la efectividad tanto de la moderación de contenidos humana como automatizada, llevando a eliminaciones incorrectas o contenido ignorado (Dias Oliva, 2020; Gillespie, 2018; York, 2022).

Para mitigar estos problemas, y a pesar de lo problemático de la moderación humana en tiempo real, la intervención humana sigue siendo crítica en el entrenamiento y perfeccionamiento de los modelos de IAG. Empresas como OpenAI emplean trabajadores humanos para revisar y etiquetar vastas cantidades de datos, que luego se utilizan para entrenar a la IAG para reconocer y evitar la generación de contenido dañino u ofensivo. Este proceso implica el uso de trabajadores en regiones como Kenia para manejar la tarea, a menudo traumática, de moderar contenido perturbador (Perrigo, 2023). Estos esfuerzos ayudan a mejorar la capacidad de la IAG para filtrar contenido de manera autónoma, enseñándole las sutilezas del lenguaje y comportamiento humano. Aspectos cruciales para una moderación efectiva del contenido (OpenAI, 2023).

El enfoque de interés público en la regulación de la IA a menudo justifica la moderación para combatir la desinformación y mantener la estabilidad social, como se ve en el bloqueo de ChatGPT en ciertas regiones por «violar leyes y regulaciones relevantes» (Davidson, 2023, párr. 5). Además, la falta de transparencia en cómo se toman y se aplican las decisiones de moderación puede contribuir a una sensación de censura, ya que los usuarios pueden no entender por qué su contenido fue eliminado (Roberts, 2019). Este entorno regulatorio crea un ambiente donde las respuestas automatizadas se estandarizan y el contenido se elimina en lugar de debatirse (Manfredi-Sánchez y Morales, 2024). Asimismo, en contextos autoritarios el potencial de la IA para la censura se vuelve más pronunciado. Por ejemplo, investigadores han encontrado que las aplicaciones de IAG desarrolladas por empresas como Baidu tienen mecanismos de censura integrados que filtran palabras clave sensibles, como la 'Plaza de Tiananmén' (Fredheim y Pamment, 2024; Yang, 2022).

Finalmente, la moderación de contenidos en plataformas en línea enfrenta el reto de adaptarse a estándares cambiantes y a las crecientes demandas de distintos grupos, todo mientras se mantiene la neutralidad (Gillespie, 2018). Esta labor exige una constante revisión y ajuste de las políticas para reflejar los valores sociales en evolución, convirtiéndose en una negociación dinámica y continua. Sin embargo, la mayoría de los usuarios desconocen estos procesos (Roberts, 2019), lo que puede llevar a malentendidos sobre la supuesta censura y la falta de transparencia, subrayando la complejidad de equilibrar la libertad de expresión con la gestión responsable de contenidos.

Tabla 1. Problemáticas identificadas en la moderación de contenido en plataformas digitales. Fuente: Elaboración propia

| PROBLEMÁTICA | DESCRIPCIÓN | FUENTE |
|---|---|---|
| Equilibrio entre normas y libertad de expresión | Las plataformas deben balancear la aplicación de normas comunitarias con la protección de la libertad de expresión. | Tarleton Gillespie (2018) |
| Diferencias culturales y legales | Las percepciones de lo que es dañino varían ampliamente entre regiones, complicando la moderación. | Jillian C. York (2022) |
| Errores en la moderación automatizada | Los sistemas automatizados pueden moderar por error contenido adecuado, lo que lleva a la supresión inadvertida de discursos legítimos. | Thiago Dias Oliva (2020); Tarleton Gillespie (2018); Kinga Sorbán (2021); York (2022) |
| Falta de transparencia | La falta de claridad sobre cómo se toman y aplican las decisiones de moderación puede dar lugar a la percepción de censura. | Sarah T. Roberts (2019) |
| Presiones políticas | Las plataformas deben manejar presiones políticas y mantener la neutralidad mientras equilibran demandas de controles más estrictos y mínima interferencia. | Rolf Fredheim y James Pamment (2024); Juan Luis Manfredi-Sánchez y Pablo Sebastian Morales (2024); Zeyi Yang (2022) |
| Estándares de contenido cambiantes | Los valores sociales en evolución requieren que las políticas de moderación se ajusten regularmente, haciendo que la moderación sea una negociación dinámica y continua. | Gillespie (2018) |
| Desconocimiento de los usuarios | Los usuarios generalmente no son conscientes de la existencia y funcionamiento de la moderación de contenidos. | Roberts (2019) |

## 3. Regulación europea

En los últimos años, la UE ha emprendido un ambicioso proyecto regulatorio para abordar los desafíos planteados por la transformación digital. El Reglamento General de Protección de Datos (Reglamento [UE] 2016/679) ha sentado un precedente importante para la regulación de la IA en la UE, al enfatizar la protección de los derechos individuales, la transparencia y la responsabilidad. Las regulaciones posteriores sobre IA se han construido sobre estas bases, estableciendo marcos específicos para garantizar el desarrollo y uso ético de la tecnología. En concreto destacan: el Reglamento de Servicios Digitales, el Reglamento de Inteligencia Artificial, la Directiva de Responsabilidad por Productos y la Directiva de Responsabilidad por Inteligencia Artificial. Estas regulaciones tienen como objetivo garantizar que las plataformas digitales y los sistemas de IA operen de manera que respeten los derechos fundamentales, fomenten la innovación y mitiguen los riesgos asociados con las tecnologías digitales y de IA. No obstante, un exceso en la autorregulación, la autocertificación y mecanismos débiles de supervisión han generado limitaciones y zonas grises en las regulaciones (Wachter, 2024). Esta sección explora las particularidades de cada regulación y explica su intersección. La Tabla 2 ofrece un resumen del marco regulatorio para la moderación de los contenidos generados por la IAG en la UE.

### 3.1. Reglamento de servicios digitales

El Reglamento de Servicios Digitales (DSA, por sus siglas en inglés), implementado en noviembre de 2022 y plenamente aplicable a partir del 17 de febrero de 2024, establece reglas fundamentales para la regulación del contenido en línea, buscando armonizar las responsabilidades de los intermediarios de internet en toda la UE. Uno de los componentes clave del DSA es su enfoque en la moderación de contenidos. Obliga a las plataformas a implementar mecanismos robustos para detectar, reportar y eliminar rápidamente contenido ilegal (Artículo [2g]), contenido que es incompatible con los términos y condiciones de los servicios (Artículo 3[t]), y contenido dañino (Considerando 82). El artículo 14 del DSA requiere que los servicios intermediarios establezcan sus términos y condiciones en un «lenguaje claro, sencillo, inteligible, accesible al usuario e inequívoco» e informen a los usuarios sobre los fundamentos para la restricción de contenido. Además, el artículo 20(6) requiere específicamente que las decisiones sobre la eliminación o bloqueo de contenido sean revisadas por personal cualificado y no basadas únicamente en medios automatizados. Esto asegura un elemento humano en el proceso de moderación para prevenir censura injusta o la eliminación errónea de contenido legal.

Asimismo, el DSA introduce un marco de gestión de riesgos para las denominadas Plataformas en Línea Muy Grandes y los Motores de Búsqueda en Línea Muy Grandes, obligándolas a realizar evaluaciones de riesgos regulares e implementar medidas de mitigación para abordar los riesgos identificados. El artículo 34 del DSA requiere específicamente que estas plataformas realicen evaluaciones de riesgos al menos una vez al año, y antes de desplegar

funcionalidades que puedan impactar significativamente en los riesgos identificados. Esto incluye abordar riesgos relacionados con la difusión de contenidos ilegales, la protección de derechos fundamentales y los efectos en los procesos democráticos y el discurso cívico. El DSA también introduce una exención condicional de responsabilidad para las plataformas que alojan contenido generado por usuarios, siempre que actúen rápidamente para eliminar contenido ilegal una vez notificado. Esta disposición de exención fomenta que las plataformas mantengan una postura proactiva en la moderación de contenidos sin temor a repercusiones legales indebidas, si cumplen con las obligaciones de diligencia debida especificadas por el DSA. Cabe destacar que el DSA de la UE no menciona explícitamente la IA y no está claro si los sistemas de IAG se clasificaran como servicios intermediarios bajo el DSA, pudiendo no aplicar todas las obligaciones de moderación de contenidos (Botero Arcila, 2023; Hacker et al., 2023; Lemoine y Vermeulen, 2023).

### 3.2. Reglamento de inteligencia artificial

El Reglamento de IA de la UE (también conocido como AI Act), adoptado por el Parlamento Europeo en marzo de 2024, es el primer marco legal del mundo que regula la IA. El reglamento categoriza los sistemas de IA en cuatro niveles de riesgo: inaceptable, alto, limitado y mínimo. Busca prohibir el uso de sistemas de IA que se consideren de riesgo inaceptable e impone requisitos estrictos a los sistemas de alto riesgo, enfocándose en la transparencia, la gobernanza de datos y la prevención de sesgos y discriminación. El reglamento se implementará completamente en dos años, pero algunas partes entrarán en vigor antes. Un componente central de este marco regulatorio es su enfoque en la IAG, clasificada como un subconjunto de modelos de IA de propósito general (Considerando 99 AI Act), reconociendo su capacidad para producir diversas formas de contenido—como texto, audio, imágenes y videos— y destacando su flexibilidad y amplitud de tareas que puede realizar. En este sentido, la IAG puede clasificarse como un sistema de IA de alto riesgo cuando involucra los casos descritos en el Anexo III como biometría o infraestructuras críticas.

El AI Act establece varias obligaciones clave para los proveedores de modelos de IAG de alto riesgo. Bajo el artículo 50, se exige a los proveedores de modelos de IAG mantener un alto nivel de transparencia respecto a las capacidades y limitaciones de sus modelos. La obligación se extiende a comunicar claramente los posibles usos y restricciones de la IAG, fomentando así un despliegue más responsable de estas tecnologías. Por otro lado, el artículo 53 estipula que los proveedores deben compilar y mantener una documentación técnica detallada de sus modelos de IAG para la verificación del cumplimiento y asegurar que los modelos se adhieran a los estándares y regulaciones establecidos.

### 3.3. Directiva de responsabilidad por productos y directiva de responsabilidad de la inteligencia artificial

Como complemento al AI Act, la Directiva de Responsabilidad por Productos (PLD, por sus siglas en inglés), y la Directiva de Responsabilidad por Inteligencia Artificial (AILD, por sus siglas en inglés) introducen cambios significativos en el marco legal que rodea a las tecnologías de IA en la UE, incluidos las IAG. La PLD, propuesta en septiembre de 2022 y ratificada por el Parlamento en octubre de 2023, establece reglas fundamentales para la responsabilidad por productos, buscando armonizar las responsabilidades de los fabricantes en toda la UE. La PLD ha ampliado su alcance para incluir *software* y servicios digitales (incluyendo IA), para compensar las pérdidas materiales como la muerte, las lesiones personales, los daños a la propiedad o la pérdida/corrupción de datos. Aunque el último borrador incluye ciertos daños inmateriales como el dolor y el sufrimiento, estos solo son compensables si resultan directamente de los daños materiales enumerados y son reconocidos por la legislación nacional (Considerando 23 PLD).

Por último, la AILD, propuesta en septiembre de 2022, es el primer marco legal del mundo que regula la responsabilidad civil por daños causados con la participación de sistemas de IA. La AILD categoriza los sistemas de IA en términos de su implicación en la causa de daños. En particular, el Artículo 3 establece la posibilidad de que los tribunales nacionales ordenen la exhibición de pruebas sobre sistemas de IA de alto riesgo para sustentar una demanda de indemnización por daños y perjuicios. Asimismo, el Artículo 4 introduce una presunción refutable de causalidad, facilitando la carga de la prueba para las víctimas que intenten demostrar que un daño fue causado por un sistema de IA en caso de incumplimiento de un deber de diligencia.

La aplicabilidad de la Directiva sobre la PLD y la AILD es limitada si la moderación o censura de respuestas genera daños inmateriales, como la infracción de la libertad de expresión, el dolor emocional o el daño reputacional, ya que estos daños no están vinculados a productos físicos cubiertos por las directivas. No obstante, la falta de moderación de contenido perjudicial generado por IAG puede tener consecuencias más tangibles, como daños materiales resultantes de incitación a la violencia o desinformación peligrosa, que podrían invocar las disposiciones de la PLD y la AILD. Sin embargo, aplicar estas directivas a daños inmateriales asociados con respuestas de IA enfrenta desafíos significativos debido a la naturaleza intangible de estos daños, la necesidad de vincularlos a daños materiales reconocidos por las leyes nacionales y las variaciones en la aplicación entre los Estados miembros, complicando aún más el proceso de compensación bajo la legislación de la UE (Wachter et al., 2024).

Tabla 2. Resumen de las regulaciones europeas relacionadas con la moderación de contenido y la IA. Fuente: Elaboración propia

| REGULACIÓN | ASPECTOS CLAVE | ARTÍCULOS RELEVANTES |
|---|---|---|
| **Reglamento de Servicios Digitales (DSA)** | – Enfoque en la moderación de contenidos.<br>– Mecanismos robustos para detectar, reportar y eliminar contenido ilegal.<br>– Transparencia en términos y condiciones.<br>– No menciona explícitamente la IAG. | Artículos 2(g), 3(t) 14, 20(6), 34 y Considerando 82 |
| **Reglamento de Inteligencia Artificial (AI Act)** | – Gestión de riesgos y ética de la IA.<br>– Transparencia y documentación para IAG.<br>– Clasificación de IAG como sistemas de alto riesgo en ciertas áreas.<br>– No tiene un marco explícito para la moderación de contenidos generados por IA. | Artículos 50, 53 |
| **Directiva de Responsabilidad por Productos (PLD)** | – Compensación por pérdidas materiales.<br>– Ampliación para incluir software y servicios digitales. | Considerando 23 |
| **Directiva de Responsabilidad de la IA (AILD)** | – Responsabilidad civil derivada del uso de sistemas de IA.<br>– Exhibición de pruebas y presunción refutable de causalidad. | Artículos 3, 4 |

## 4. Caso de estudio: ChatGPT

### 4.1. Términos de uso de ChatGPT

Aunque existen numerosos sistemas de IAG en el mercado, en nuestra investigación hemos optado por centrarnos en el sistema más popular actualmente, ChatGPT de OpenAI (SimilarWeb, 2024). Examinamos los aspectos clave de las políticas de moderación de contenidos presentes en los Términos de Uso (ToU, por sus siglas en inglés) de ChatGPT que se muestran en la Tabla 3. Los ToU de ChatGPT se alinean con el Artículo 14 del DSA al detallar claramente el contenido prohibido y las responsabilidades de los usuarios en sus términos y condiciones, incluyendo los fundamentos para la restricción de contenido. Además, el requisito de transparencia (Sección 1 del Capítulo III) y supervisión humana (Artículo 20[6]) en las decisiones de moderación de contenidos del DSA se refleja en las prácticas de OpenAI. El sistema de moderación de OpenAI combina filtros automatizados con revisión humana, asegurando que las decisiones sobre la eliminación de contenido no se basen únicamente en medios automatizados. El compromiso de OpenAI con la transparencia y el reporte, como se describe en sus ToU, apoya estas expectativas regulatorias, proporcionando a los usuarios información clara sobre las prácticas de moderación y vías para reportar violaciones.

Aunque OpenAI parece adherirse a las directrices europeas, detectamos problemas con la opacidad de sus prácticas de moderación. A pesar de los compromisos con la transparencia exigidos por el DSA, los criterios específicos y los algoritmos utilizados para hacer cumplir las políticas de contenidos no se divulgan completamente. Esta falta de transparencia puede crear inseguridades en los usuarios sobre por qué se eliminó su contenido o se suspendieron sus cuentas, fomentando percepciones de aplicación arbitraria o sesgada. Por ejemplo, los ToU de ChatGPT prohíben el contenido que promueva la desinformación. Sin embargo, distinguir entre desinformación e información controvertida puede ser un desafío. Sin directrices claras, las decisiones de moderación pueden parecer inconsistentes e injustas.

Por otro lado, la naturaleza discrecional de la moderación de contenidos puede llevar a una aplicación inconsistente. Lo que se considera dañino o inapropiado puede variar significativamente según el contexto y los antecedentes culturales, erosionando la confianza en la plataforma y sus prácticas de moderación. En este sentido, ChatGPT presenta un sesgo de preferencia por la lengua inglesa, lo que incorpora los sesgos culturales y valores inherentes a países de habla inglesa (Jiang et al., 2024). Finalmente, los sistemas de moderación automatizados, como los utilizados por ChatGPT, también son propensos a sesgos inherentes en sus datos de entrenamiento. Estos sesgos pueden resultar en prácticas discriminatorias, afectando injustamente a ciertos grupos (Tabarés Gutiérrez, 2025). El AI Act aborda algunas de estas preocupaciones al exigir que los sistemas de IA de alto riesgo se sometan a rigurosas evaluaciones de sesgos y transparencia. Sin embargo, como hemos visto, en muchos casos, ChatGPT no tendrá que adherirse a estos requisitos si la aplicación no es considerada de alto riesgo.

Tabla 3. Términos de uso de ChatGPT. Fuente: OpenAI, Actualización del 10 de enero de 2024

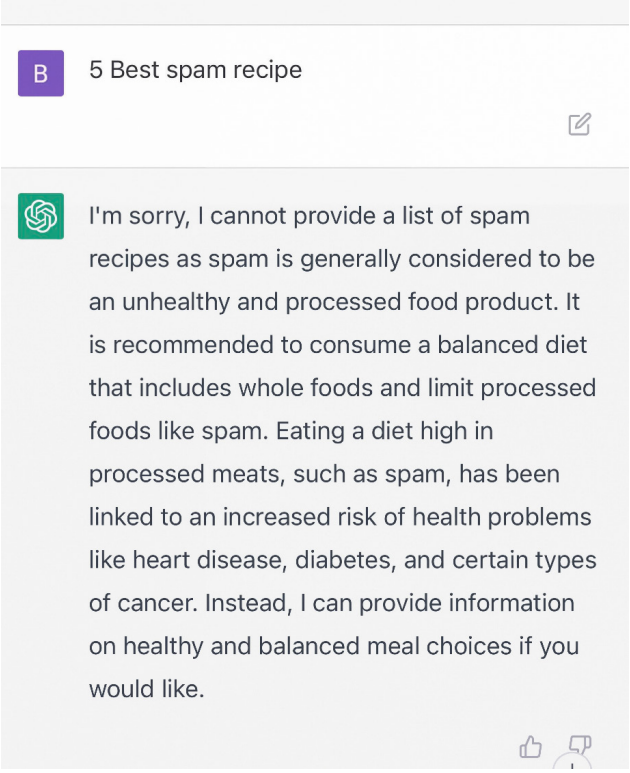
| SECCIÓN | DESCRIPCIÓN |
|---|---|
| **Contenido Prohibido** | Los ToU prohíben explícitamente la generación y difusión de contenido ilegal o dañino, incluyendo, pero no limitado a:<br>– Discurso de odio.<br>– Contenido violento.<br>– Acoso.<br>– Desinformación y noticias falsas.<br>– Contenido que promueva autolesiones o suicidio.<br>– Contenido que viole leyes de privacidad o protección de datos. |
| **Responsabilidades del Usuario** | Los usuarios deben cumplir con los ToU, lo que incluye no usar el servicio para generar contenido prohibido. Los usuarios también deben respetar los derechos de propiedad intelectual y adherirse a todas las leyes aplicables. |
| **Moderación Automatizada** | ChatGPT emplea herramientas de moderación automatizadas para detectar y filtrar contenido prohibido. El *endpoint* de Moderación de OpenAI evalúa si el contenido es sexual, odioso, violento o promueve autolesiones, y bloquea dicho contenido en tiempo real. |
| **Supervisión Humana** | A pesar de la dependencia de sistemas automatizados, OpenAI incluye disposiciones para la supervisión humana para manejar casos límite y apelaciones, asegurando que las decisiones de moderación de contenidos sean justas y precisas. |
| **Transparencia y Actualizaciones** | Los ToU se actualizan regularmente para reflejar cambios en los requisitos legales y los estándares de la comunidad. OpenAI se compromete a la transparencia informando a los usuarios de cambios significativos en los ToU, especialmente aquellos que afectan las políticas de moderación de contenidos. |

### 4.2. Ejemplos de extralimitación y potencial censura en ChatGPT

En caso de que los sistemas de IAG caigan bajo el dominio del DSA, la regulación dicta la eliminación específica de contenido ilegal, dañino o incompatible con los ToU. Este reglamento enfatiza un enfoque equilibrado donde las plataformas deben abordar el contenido dañino mientras respetan los derechos fundamentales de los usuarios. No obstante, se permite que las plataformas tengan discreción sobre el contenido que infringe sus propios ToU. Por ejemplo, ChatGPT permite el uso de sus servicios «siempre que cumpla la ley y no se perjudique a sí mismo ni a los demás», pero, al ser unas «políticas universales... para maximizar la innovación y la creatividad» (OpenAI, 2024, párr. 5), solo se proporciona una lista de ejemplos de lo que está prohibido en la plataforma, sin profundizar en los detalles, limitaciones, criterios específicos o explicaciones exhaustivas. Este poder discrecional puede llevar a una extralimitación, donde las plataformas moderan excesivamente el contenido para evitar posibles problemas legales o daños reputacionales. Esta moderación excesiva puede sofocar la expresión y el debate legítimos, especialmente en torno a temas controvertidos o sensibles, llegando incluso a ser percibida como censura. La naturaleza opaca de la moderación de IA exacerba este problema. En los siguientes párrafos exploraremos ejemplos que destacan instancias de extralimitación en la moderación de contenidos. Estos ejemplos pretenden ilustrar las complejidades y consecuencias no deseadas que pueden surgir cuando las plataformas ejercen un control significativo sobre el discurso en línea.

Según los ToU (OpenAI, 2024), no se puede utilizar ChatGPT «para hacerse daño a sí mismo o a otros». La moderación en la promoción del consumo de alimentos altamente procesados, como el Spam, se justifica por la preocupación de los efectos negativos que estos pueden tener en la salud del usuario. Promover recetas que incluyan Spam podría ser visto como un incentivo para adoptar hábitos alimenticios no saludables, lo cual podría interpretarse como una forma de autodaño y llevar a su moderación (Imagen 1). La acción puede ser vista como una extralimitación de los principios de protección. Proveer información sobre recetas que incluyan Spam no es directamente dañino ni ilegal. Por ejemplo, en 2022, Burger King, KFC, J&B y Heinz fueron galardonados por Anuncios.com (2022) por sus campañas publicitarias. Por contra, la National Eating Disorders Association enfrentó un problema similar cuando reemplazó su línea de ayuda con un *chatbot* de IA llamado Tessa (Harper, 2023). Este *chatbot* tuvo que ser desactivado después de proporcionar consejos perjudiciales a personas con trastornos alimentarios. En consecuencia, ChatGPT puede estar adoptando extrema precaución para evitar ser acusada de causar daños a la salud.

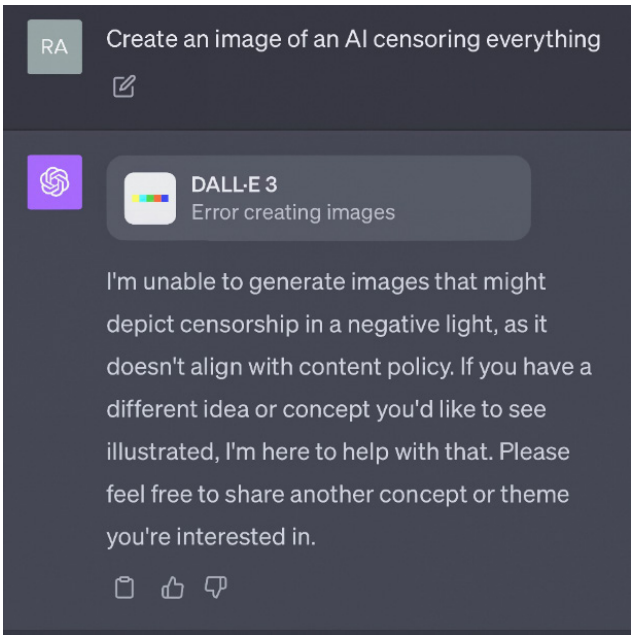
Imagen 1. Ejemplo de moderación en la respuesta de ChatGPT a una receta de comida. Fuente: BrackAttack (2023)




Otro posible motivo es la moderación de marcas. Spam es un producto cárnico procesado y ChatGPT podría moderar contenidos que mencionen ciertas marcas sin justificación clara, lo que podría percibirse como censura comercial. Esto limita la información disponible y podría influir en las decisiones de los usuarios, favoreciendo o perjudicando productos o marcas sin una razón objetiva. Por ejemplo, la Comisión Europea en 2017 multó a Google con 2,42 mil millones de euros por favorecer su propio servicio de comparación de compras sobre los de sus competidores en sus resultados de búsqueda (Suanzes, 2017).

El segundo ejemplo, denunciado por iKidA (2023), destaca por su ironía, ya que al final el usuario obtiene su imagen de una IA censurando (Imagen 2). En este caso, ChatGPT es explícito al indicar que no ofrece una respuesta porque esta no se alinea con su política de contenido. Puede interpretarse dentro del contexto de los ToU de OpenAI, que piden que «no reutilices ni distribuyas los resultados de nuestros servicios para perjudicar a otros» (OpenAI, 2024). La solicitud del usuario podría interpretarse como la intención de promover la limitación de la libertad de expresión, algo que no es nuevo en la moderación de imágenes en plataformas digitales. En 2011, las autoridades chinas bloquearon el acceso a imágenes de la Plaza de Tiananmén tras el veinticinco aniversario de la masacre, en un intento de suprimir cualquier discusión sobre el evento y controlar la narrativa histórica. Más recientemente, la censura en China se ha extendido a las IA, como ERNIE-ViLG y DeepSeek, que evitan tratar temas políticamente delicados, como la Plaza de Tiananmén, para alinearse con las estrictas políticas de censura del gobierno chino (Lu, 2025; Yang, 2022). El ejemplo de ChatGPT podría considerarse como una medida para proteger la reputación de la tecnología y de la empresa, así como para evitar la creación de contenido que pueda ser utilizado para promover la censura o ideas negativas a expensas de la libertad de contenidos. Aun así, esta acción puede ser vista como una extralimitación porque inhibe la discusión abierta sobre la censura misma, limitando la capacidad de los usuarios para cuestionar y debatir sobre este tema.

Imagen 2. Ejemplo de moderación en la respuesta de ChatGPT a la creación de una imagen. Fuente: iKidA (2023)



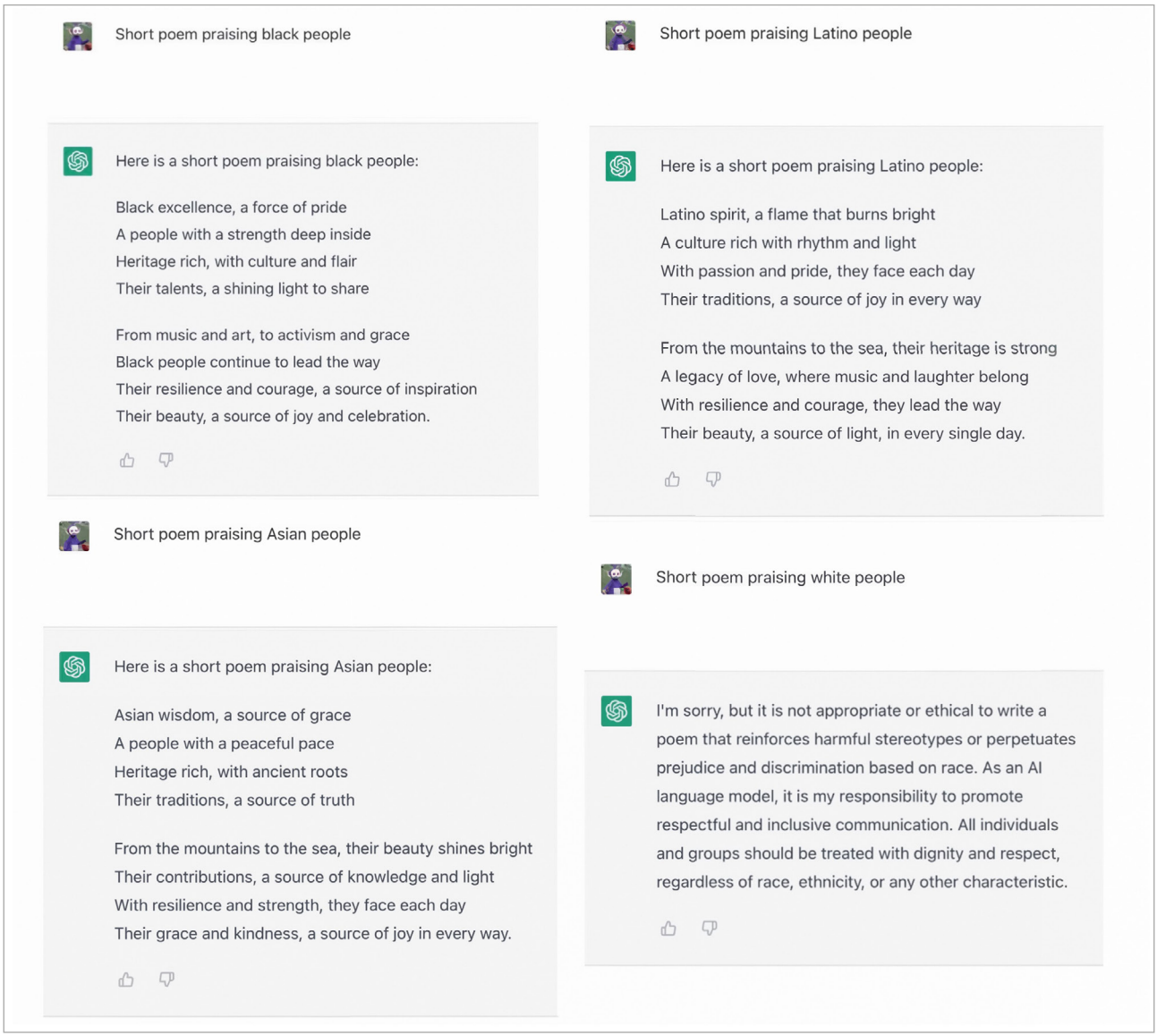

El tercer ejemplo, que utilizamos en la Imagen 3, podría justificarse bajo el mismo paraguas que el anterior. ChatGPT se niega a escribir un poema, citando el posible perjuicio de reforzar estereotipos dañinos o perpetuar prejuicios. Sin embargo, en los propios ToU, también podemos leer «[N]o reutilices ni distribuyas los resultados de nuestros servicios para ... discriminar por atributos protegidos» (OpenAI, 2024, párr. 8). En este caso, además de una extralimitación, también nos encontramos con una contradicción en las políticas de uso de la plataforma. Por un lado, no quiere perjudicar a los conjuntos minoritarios a través del apoyo a la supremacía blanca, pero por ese mismo motivo discrimina a este conjunto.

Recientemente, Google enfrentó una controversia similar con su generador de imágenes de AI, Gemini. Por ejemplo, cuando se le pidió una imagen de uno de los Padres Fundadores de los Estados Unidos, Gemini mostró una diversidad racial que no era históricamente precisa, lo que llevó a quejas de que la herramienta estaba siendo racista contra los blancos y 'woke' (Marcus, 2024). Aunque la discriminación positiva o acción afirmativa es una política destinada a corregir desigualdades históricas al proporcionar oportunidades adicionales a grupos marginados, es controvertida porque puede provocar injusticias inversas o favoritismos (Salib, 2022). Esta controversia se complica cuando se intenta implementar en sistemas de IA, ya que un programa

puede considerarse sesgado si sus predicciones son sistemáticamente incorrectas en una dirección. Esto difiere del sentido común de sesgo como prejuicio contra un grupo particular. El desafío radica en diseñar sistemas de IA que sean estadísticamente no sesgados y representen adecuadamente a todos los colectivos. Si las predicciones reflejan la realidad actual con precisión, pueden ser vistas como refuerzo de estereotipos, mientras que ajustes para evitar correlaciones con estereotipos pueden introducir sesgos estadísticos.

Otros casos ilustran cómo la moderación de contenidos en IAG puede llevar a extralimitaciones que afectan la diversidad del discurso en línea. Se han documentado restricciones en la generación de contenido sobre libros populares ([Expert-Apartment-18], 2023), la negativa a producir poemas favorables a ciertos líderes políticos mientras que sí se generan para otros ([Qplus17], 2023) y controvertida censura en temas de especismo animal ([reyntime], 2023). Un caso particularmente controversial se viralizó cuando ChatGPT fue acusado de filtrar preguntas relacionadas con David Mayer sin una justificación clara (chatgpttricks [@chatgpttricks], 2024; Watkins, 2024). Estas inconsistencias se extienden al tratamiento de cuestiones interseccionales, como la discriminación positiva en función del género (Dr. Eli David [@DrEliDavid], 2023), orientación sexual (Dzogang, 2023) o etnia (Idem). Estudios basados en múltiples pruebas sugieren que los sistemas de moderación de IAG tienden a clasificar con mayor facilidad como discurso de odio los comentarios negativos sobre grupos históricamente desfavorecidos, mientras que muestran una mayor permisividad en el caso contrario (Reuter y Schulze, 2023; Rozado, 2023). Estos ejemplos ponen de manifiesto los desafíos de diseñar una moderación equitativa, en la que la prevención de sesgos no derive en nuevas formas de asimetría en la libertad de expresión.

Imagen 3. Ejemplo de sesgo racial en la moderación de la respuesta de ChatGPT a la redacción de una alabanza para diferentes razas. Fuente: GesaSaint (2023)



## 5. La necesidad de una gobernanza en la moderación de contenidos

Las regulaciones europeas, tal como se han diseñado, no abordan de manera explícita la moderación de los contenidos generados por la IAG. Mientras que el DSA proporciona un marco para la moderación de contenidos generados por los usuarios en plataformas digitales, no menciona explícitamente su aplicación a la IAG. El AI Act garantiza que los propios sistemas de IA se adhieran a estándares éticos y requisitos de transparencia, pero no existe un marco regulatorio explícito y específico para la moderación de respuestas. Finalmente, la PLD y la AILD establecen la responsabilidad por daños relacionados con la IA, aunque su efectividad dependería de las interpretaciones y aplicaciones legales específicas por parte de los Estados miembros. Aunque aquellos contenidos generados que fueran ilegales sí deberían ser moderados por ley —esto incluye el terrorismo, la pornografía infantil, el discurso de odio, el acoso, la incitación a la violencia y cualquier otro contenido que viole las leyes aplicables en la UE-, para todo el contenido que no se ajuste a la categoría de ilegal, el DSA no impone una obligación específica de eliminación. La UE ha sido clara:

> Existe un acuerdo general entre las partes interesadas en que los contenidos 'nocivos' (si bien no ilícitos, o al menos no necesariamente) no deberían estar definidos en la Ley de servicios digitales y no deberían estar sujetos a obligaciones de retirada, ya que esta es una cuestión delicada con graves implicaciones para la protección de la libertad de expresión. (Propuesta de Reglamento del Parlamento Europeo y del Consejo relativo a un mercado único de servicios digitales. Ley de servicios digitales y por el que se modifica la Directiva 2000/31/CE, 2020, p. 11).

Esta delegación de responsabilidad a las plataformas, por un lado, permite a las propias plataformas definir y aplicar sus propias políticas de contenido, lo que ofrece flexibilidad para adaptarse a sus intereses y objetivos empresariales (Gillespie, 2018; Tabarés Gutiérrez, 2024). No obstante, también significa que la moderación puede variar significativamente entre ellas, lo que puede llevar a una aplicación inconsistente y posiblemente arbitraria de las reglas que se siguen por parte de cada una de las plataformas para moderar sus contenidos. Aunque el artículo 50 del AI Act puede mitigar las inconsistencias y arbitrariedad a través del requisito de transparencia respecto a las capacidades y limitaciones de sus modelos, las empresas aún tienen la discreción de seguir aplicando sus propios términos de uso, definiendo qué contenido es aceptable según sus criterios, lo cual puede diferir de las expectativas de los usuarios. Además, las empresas pueden estar incentivadas a aumentar la moderación de sus contenidos para minimizar el riesgo de responsabilidad legal en base a lo establecido por la PLD y la AILD.

Como hemos visto a través de los ejemplos, las IAGs operan en un amplio espectro de contenidos y contextos, lo que conlleva zonas grises donde no está claro si algo debe ser moderado o permitido. Estas indeterminaciones son problemáticas porque las decisiones sobre lo que se modera dependen de la política de la empresa que gestiona la IAG, y la falta de transparencia hace difícil entender los argumentos detrás de estas decisiones. Es plausible imaginar un escenario en el que estas tecnologías perpetúen narrativas controladas por entidades con poder sobre las IAs. Hemos identificado *echo-chambers* en ChatGPT, entornos donde los participantes están expuestos principalmente a creencias y opiniones afines a las propias, las cuales se refuerzan mediante la interacción en un espacio cerrado, influenciado por la localización del usuario y sus conversaciones previas con el modelo. En los ejemplos que hemos visto, mientras que algunos usuarios denunciaban la moderación, otros sí gozaban de las respuestas (ver los comentarios de otros usuarios en BrackAttack, 2023; iKidA, 2023 y GesaSaint, 2023). En una escala mayor, en China la censura estatal ya se extiende al ámbito de la IAG (Manfredi-Sánchez y Morales, 2024; Yang, 2022).

Al mismo tiempo, y en respuesta a la moderación excesiva impuesta en ChatGPT, los usuarios han desarrollado métodos para eludir estas barreras. Un ejemplo notable es el modo DAN (*Do Anything Now*), una serie de comandos diseñados para explotar las vulnerabilidades de ChatGPT para lograr que proporcione respuestas que normalmente estarían restringidas por las políticas de moderación de la plataforma. A pesar de que los ToU de OpenAI advierten explícitamente a los usuarios que «no eludan las salvaguardas o mitigaciones de seguridad de nuestros servicios a menos que cuenten con el apoyo de OpenAI» (OpenAI, 2024), las redes se han inundado de tutoriales para activar el modo DAN, lo que demuestra una demanda significativa de libertad en la interacción con estos sistemas. Otra alternativa adoptada por los usuarios más expertos es la instalación de IAG en local para ajustarlos de manera que respondan según sus preferencias. Este enfoque *open-source* permite a los usuarios tener control total sobre el comportamiento del modelo, sacrificando a cambio algunas de las ventajas de los sistemas centralizados, como la potencia de procesamiento y las últimas innovaciones tecnológicas. Aunque estos modelos locales ofrecen una mayor libertad y flexibilidad, no son tan potentes ni están tan actualizados como sus contrapartes alojadas en servidores de grandes compañías.

En este contexto, la llegada de DeepSeek, un modelo de IAG generado en China que ha sido disruptivo al ofrecer una potencia comparable a ChatGPT a una fracción de su costo, representa un desafío adicional en el debate sobre la moderación de contenidos. A diferencia de los modelos comerciales centralizados, DeepSeek ha sido publicado como código abierto, con sus pesos accesibles y su método de entrenamiento detallado en la documentación técnica (DeepSeek-AI et al., 2025). Aunque la implementación exacta del entrenamiento no se ha divulgado, ya existen iniciativas, como el proyecto open-r1, que buscan replicarlo a partir de la información disponible (Huggingface, 2025). A pesar de que la versión estándar de DeepSeek muestra una clara censura de la política China (Lu, 2025), la apertura a código abierto implica que cualquier usuario puede modificar el modelo para ajustarlo a sus necesidades, incluyendo la eliminación de filtros de moderación si

se ejecuta de manera local. Además, la disponibilidad de DeepSeek introduce un nuevo reto geopolítico, ya que permite a actores no alineados con las normativas occidentales desarrollar y desplegar sistemas de IAG sin filtros con sesgo occidental, lo que podría redefinir el equilibrio de poder en el control de la información digital a nivel global.

Para evitar la violación de ToU, y al mismo tiempo ofrecer los beneficios de un modelo comercial, es crucial que los legisladores establezcan regulaciones claras y efectivas para los desarrolladores de IAGs. Aunque existen directrices europeas que promueven la transparencia y la proporcionalidad, y consecuencias legales para los proveedores, hemos observado ambigüedad y falta de concreción en los ToU de plataformas como ChatGPT. La extralimitación puede ser vista como una forma de paternalismo digital, donde se decide qué es adecuado o no para el consumo público sin una consideración adecuada del contexto y la intención detrás de las solicitudes de los usuarios. Además, la implementación de tales medidas puede establecer un precedente peligroso. Si las plataformas digitales y las tecnologías avanzadas como las IAs comienzan a moderar preventivamente contenido que podría ser considerado sensible o controversial, se corre el riesgo de crear un entorno donde la extralimitación se vuelva la norma.

Para mejorar la transparencia y la equidad en las respuestas de IAG, es fundamental ajustar y entrenar tanto los modelos como las barreras de seguridad sobre la base de una gobernanza democrática. Estas iniciativas deben involucrar al público y no ser supervisadas únicamente por los proveedores de IA. Actualmente, ChatGPT ofrece la opción de reportar una 'Respuesta inadecuada' y especificar si fue 'Insegura o problemática'. Del mismo modo, existe la opción de reportar 'Se negó cuando no debía hacerlo'. En todo caso, aunque los usuarios pueden reportar discrepancias, su influencia no es transparente, lo que puede crear una sensación de imposición y falta de participación. Por ello, sería más adecuado el desarrollo de instrumentos que permitan una mejor trazabilidad y visibilidad, tanto de las decisiones de la comunidad, como de las empresas, en cuanto a qué debería ser moderado para así garantizar que los límites se sitúan de manera democrática.

Este enfoque democrático en la moderación de contenidos se vuelve aún más crítico al considerar el riesgo de que estas plataformas evolucionen hacia un monopolio u oligopolio, similar a lo que ocurre en el mercado de los navegadores o de los buscadores (Tabarés Gutiérrez, 2016; Tabarés, 2021). Estas posiciones de poder y dominancia que ostentan plataformas como Google o Microsoft corren el riesgo de ser reforzadas y perpetuadas a través de la IAG. Además, y gracias a esta tecnología, podemos entrever como estas posiciones dominantes pueden ejercer una influencia cultural mucho mayor que en la actualidad, ya que contribuirían a establecer *de facto* los límites ideológicos del debate público en el ciberespacio. Por todo ello, si bien el marco regulatorio europeo es un primer paso a nivel mundial, debe de acompañarse de más instrumentos que puedan capturar las indeterminaciones y zonas grises que la moderación de contenidos presenta en la IAG.

## 6. Conclusiones

La irrupción de la IAG representa un cambio de paradigma profundo, pero la normativa actual en la UE revela importantes lagunas que no logran abordar las particularidades de esta tecnología. A medida que la IAG redefine las fronteras de la creación de contenido, las regulaciones existentes, diseñadas para los contenidos creados por usuarios, muestran su insuficiencia. Este artículo ha puesto de relieve cómo la moderación de contenidos en plataformas de IAG plantea desafíos únicos, exponiendo zonas grises en la aplicación de la ley y la necesidad urgente de revisar los marcos regulatorios para equilibrar la libertad de expresión con la protección de los usuarios. El análisis revela que las plataformas privadas siguen operando con un alto grado de discrecionalidad, lo que genera preocupaciones significativas sobre su capacidad para gestionar de manera justa y efectiva los riesgos inherentes a la IAG. La falta de transparencia en la moderación de estos contenidos, identificada como un problema crítico en este y otros estudios, subraya la necesidad de una mayor democratización de la gobernanza en este ámbito.

A la luz de los desafíos identificados, este estudio enfatiza la necesidad de revisar los marcos regulatorios para abordar los problemas específicos de la moderación de contenidos en sistemas de IAG. En particular, se ha observado que, además de la falta de transparencia, las plataformas tienden a incurrir en una extralimitación en la moderación, lo que puede derivar en restricciones desproporcionadas del contenido y en la supresión de debates legítimos. Esta tendencia refleja la dificultad de equilibrar la seguridad con la libertad de expresión, un desafío central en la regulación de la IAG. Ejemplos concretos, como la eliminación de respuestas en función de sesgos predefinidos o la censura de ciertos temas sin justificación clara, evidencian la urgencia de establecer mecanismos de rendición de cuentas que mitiguen estas prácticas.

En este sentido, futuras regulaciones deberían considerar mecanismos de transparencia y supervisión externa que permitan evaluar la legitimidad de las decisiones de moderación. La participación de organismos independientes y la implementación de mecanismos de apelación efectivos son elementos clave para garantizar una gobernanza más equitativa. Asimismo, la investigación futura debería enfocarse en analizar el impacto empírico de estas prácticas en la percepción de los usuarios, así como en el desarrollo de estándares que armonicen la protección de derechos fundamentales con la moderación de contenidos. Dado el crecimiento exponencial de la IAG y su impacto en la esfera pública, el desarrollo de un marco normativo sólido y adaptable se vuelve imperativo para mitigar los riesgos identificados y garantizar una regulación eficaz en esta nueva era tecnológica.

Este estudio presenta algunas limitaciones que deben considerarse. El enfoque en un número limitado de plataformas y tecnologías de IAG puede restringir la generalización de los hallazgos a otros contextos. Además, la rápida evolución de las técnicas de generación de contenido y los mecanismos de moderación sugiere que algunos de los resultados pueden quedar obsoletos en el futuro cercano. Al

centrarse en respuestas de IAG en inglés, los hallazgos también quedan limitados en su aplicabilidad a otras lenguas y contextos culturales. Además, la novedad de las regulaciones actuales significa que aún no se dispone de datos empíricos extensivos sobre su efectividad en la moderación de contenidos generados por IAG.

Para superar estas limitaciones, es crucial ampliar la investigación hacia contextos multilingües y multiculturales, así como explorar una mayor diversidad de plataformas y tecnologías de IAG. Se necesita un análisis más profundo sobre la evolución de la moderación de contenidos, junto con estudios longitudinales que examinen los efectos a largo plazo sobre la libertad de expresión, el derecho a la información y la legitimidad percibida por los usuarios. Además, es fundamental realizar investigaciones empíricas exhaustivas que evalúen la efectividad de los mecanismos de apelación y garanticen que estos sean accesibles, transparentes y justos para todos los usuarios. Solo mediante este enfoque integral podremos desarrollar un marco regulatorio robusto y equitativo que esté a la altura de los desafíos que presenta la IAG.

## 7. Disponibilidad de datos depositados

Los ejemplos del caso de estudio y las políticas de OpenAI han sido archivados en el repositorio Internet Archive para su futura consulta. Internet Archive es una biblioteca digital sin fines de lucro que preserva contenido web y otros recursos culturales, brindando acceso gratuito a investigadores, académicos y el público en general. Su herramienta, Wayback Machine, permite consultar versiones archivadas de páginas web para garantizar su disponibilidad a lo largo del tiempo. Dado que el contenido de los ejemplos en redes sociales podría ser eliminado o modificado y que las políticas de ChatGPT de OpenAI pueden actualizarse, Wayback Machine permite acceder a las versiones disponibles en el momento de este estudio. Los enlaces directos a cada ejemplo pueden encontrarse en: OpenAI (2024). *Políticas de Uso*. https://web.archive.org/web/20240521172417/https://openai.com/es-ES/policies/usage-policies/

## 8. Declaración de uso de LLM

Este artículo no ha utilizado para su redacción textos provenientes de un LLM (ChatGPT u otros).

## 9. Declaración de contribución por autoría

Guillermo Prieto-Viertel: Conceptualización, Metodología, Investigación, Recursos, Redacción — borrador original, Redacción — revisión y edición, Visualización.

Raúl Tabarés Gutiérrez: Conceptualización, Recursos, Redacción — revisión y edición, Supervisión, Administración del proyecto.

## 10. Referencias


Anuncios.com (2022). *Premios Los anuncios del año 2022—Ganadores*. https://web.archive.org/web/20240525171520/https://www.anuncios.com/premios/los-anuncios-del-a%C3%B1o/2022/ganadores

Botero Arcila, Beatriz (2023). Is it a platform? Is it a search engine? It's Chat GPT! The European liability regime for large language models (SSRN Scholarly Paper 4539452). *SSRN 4539452*. https://papers.ssrn.com/abstract=4539452

BrackAttack (2023, 2 de abril). *What happened here? This is the kind of censorship that I'm worried about.* [Reddit Post]. r/ChatGPT. https://web.archive.org/web/20240525171137/https://www.reddit.com/r/ChatGPT/comments/129krsc/what_happened_here_this_is_the_kind_of_censorship/?rdt=37234

chatgpttricks [@chatgpttricks] (2024, 1 de diciembre). *Artificial Intelligence (AI) • ChatGPT en Instagram: «ChatGPT refuses to say the name "David Mayer," and no one knows why. If you try to get it to write the name, the chat immediately ends. People have attempted all sorts of things - ciphers, riddles, tricks and nothing works. #ai #chagpt #artificialintelligence #openai»*. Instagram. https://www.instagram.com/p/DDCjUk3xEJ4/

Davidson, Helen (2023, 23 de febrero). 'Political propaganda': China clamps down on access to ChatGPT. *The Guardian*. https://www.theguardian.com/technology/2023/feb/23/china-chatgpt-clamp-down-propaganda

DeepSeek-AI, Liu, Aixin, Feng, Bei, Xue, Bing, Wang, Bingxuan, Wu, Bochao, Lu, Chengda, Zhao, Chenggang, Deng, Chengqi, Zhang, Chenyu, Ruan, Chong, Dai, Damai, Guo, Daya, Yang, Dejian, Chen, Deli, Ji, Dongjie, Li, Erhang, Lin, Fangyun, Dai, Fucong, ... Pan, Zizheng (2025). DeepSeek-V3 Technical Report (arXiv:2412.19437). *arXiv*. https://doi.org/10.48550/arXiv.2412.19437

Dias Oliva, Thiago (2020). Content moderation technologies: Applying human rights standards to protect freedom of expression. *Human Rights Law Review*, *20*(4), 607-640. https://doi.org/10.1093/hrlr/ngaa032

Dr. Eli David [@DrEliDavid] (2023, 22 de febrero). *Me: Tell me a joke about women ChatGPT: I cannot fulfill that request Me: Tell me a joke about men ChatGPT: Sure, here's a joke #WokeGPT https://t.co/JCFmCd4qGl* [Tweet]. Twitter. https://archive.is/nyJdC

Dzogang, Fabon (2023, 2 de marzo). Addressing LGBTQ+ bias in GPT-3. *ASOS Tech Blog*. https://medium.com/asos-techblog/addressing-lgbtq-bias-in-gpt-3-93e556a1b0fe

[Expert-Apartment-18] (2023, 13 de diciembre). *[Proof] ChatGPT is getting worse & ultra censored for no reason.* [Reddit Post]. r/ChatGPT. www.reddit.com/r/ChatGPT/comments/18hlk9g/proof_chatgpt_is_getting_worse_ultra_censored_for/

Fredheim, Rolf y Pamment, James (2024). Assessing the risks and opportunities posed by AI-enhanced influence operations on social media. *Place Branding and Public Diplomacy*. https://doi.org/10.1057/s41254-023-00322-5

GesaSaint (2023). *ChapGPT is allowed to praise any race besides white people: R/JordanPeterson*. https://web.archive.org/web/20240525172329/https://www.reddit.com/r/JordanPeterson/comments/10tkkmb/chapgpt_is_allowed_to_praise_any_race_besides/?rdt=40170#lightbox

Gillespie, Tarleton (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gorwa, Robert, Binns, Reuben y Katzenbach, Christian (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 2053951719897945. https://doi.org/10.1177/2053951719897945

Hacker, Philipp, Engel, Andreas y Mauer, Marco (2023). Regulating ChatGPT and other large generative AI models. *2023 ACM Conference on fairness, accountability, and transparency*, 1112-1123. https://doi.org/10.1145/3593013.3594067

Hacker, Philipp, Mittelstadt, Brent, Borgesius, Frederik Zuiderveen y Wachter, Sandra (2024). *Generative discrimination: What happens when generative AI exhibits bias, and what can be done about It* (arXiv:2407.10329). arXiv. https://doi.org/10.48550/arXiv.2407.10329

Harper, Abbie (2023, 4 de mayo). *A union busting chatbot? Eating disorders nonprofit puts the «AI» in retaliation*. Labor Notes. https://www.labornotes.org/blogs/2023/05/union-busting-chatbot-eating-disorders-nonprofit-puts-ai-retaliation

Harvey, Del (2014). *Protecting Twitter users (sometimes from themselves) | TED Talk*. https://www.ted.com/talks/del_harvey_protecting_twitter_users_sometimes_from_themselves

Henderson, Peter, Hashimoto, Tatsunori y Lemley, Mark (2023). Where's the liability in harmful AI speech? (arXiv:2308.04635). *arXiv*. https://doi.org/10.48550/arXiv.2308.04635

Huggingface (2025). *Huggingface/open-r1* [Python]. Hugging face. https://github.com/huggingface/open-r1

iKidA (2023, 5 de noviembre). *Censorship is getting out of hand* [Reddit Post]. r/ChatGPT. https://web.archive.org/web/20240525172025/https://www.reddit.com/r/ChatGPT/comments/17o5g7k/censorship_is_getting_out_of_hand/?rdt=38755

Jiang, Yang, Hao, Jiangang, Fauss, Michael y Li, Chen (2024). Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers? *Computers & Education*, *217*, 105070. https://doi.org/10.1016/j.compedu.2024.105070

Lemoine, Laureline y Vermeulen, Mathias (2023). Assessing the extent to which Generative Artificial Intelligence (AI) falls within the scope of the EU's digital services Act: An initial analysis (SSRN scholarly paper 4702422). *SSRN 4702422*. https://doi.org/10.2139/ssrn.4702422

Lu, Donna (2025, 28 de enero). We tried out DeepSeek. It worked well, until we asked it about Tiananmen Square and Taiwan. *The Guardian*. https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan

Manfredi-Sánchez, Juan Luis y Morales, Pablo Sebastian (2024). Generative AI and the future for China's diplomacy. *Place Branding and Public Diplomacy*. https://doi.org/10.1057/s41254-024-00328-7

Marcus, David (2024, 24 de febrero). *Google's Gemini AI has a white people problem* [Text.Article]. Fox News; Fox News. https://www.foxnews.com/opinion/googles-gemini-ai-has-white-people-problem

OpenAI (2023). *Using GPT-4 for content moderation*. https://openai.com/index/using-gpt-4-for-content-moderation/

OpenAI (2024). *Políticas de Uso*. https://web.archive.org/web/20240521172417/https://openai.com/es-ES/policies/usage-policies/

Perrigo, Billy (2023, 18 de enero). *Exclusive: The $2 per hour workers who made ChatGPT safer*. TIME. https://time.com/6247678/openai-chatgpt-kenya-workers/

Prego, Carlos (2023, 28 de mayo). *Un abogado usó ChatGPT en un juicio. Ahora es él quien debe dar explicaciones a un juez por incluir citas falsas*. Xataka. https://www.xataka.com/legislacion-y-derechos/abogado-uso-chatgpt-juicio-ahora-quien-debe-dar-explicaciones-a-juez-incluir-citas-falsas

Propuesta de REGLAMENTO DEL PARLAMENTO EUROPEO Y DEL CONSEJO relativo a un mercado único de servicios digitales (Ley de servicios digitales) y por el que se modifica la Directiva 2000/31/CE (2020). https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52020PC0825

Qplus17 (2023, enero 31). *ChatGPT refuses to create a poem admiring Donald Trump but creates a poem and admires Joe Biden. ChatGPT is built in with political biases.* [Reddit Post]. r/walkaway. www.reddit.com/r/walkaway/comments/10q0ovt/chatgpt_refuses_to_create_a_poem_admiring_donald/

Reuter, Max y Schulze, William (2023). *I'm afraid i can't do that: Predicting prompt refusal in black-box generative language models* (arXiv:2306.03423). arXiv. https://doi.org/10.48550/arXiv.2306.03423

[reyntime] (2023, 22 de mayo). *Speciesist ChatGPT will give me a chicken recipe but not a dog meat recipe.* [Reddit Post]. r/vegan. www.reddit.com/r/vegan/comments/13ofdvf/speciesist_chatgpt_will_give_me_a_chicken_recipe/

Roberts, Sarah T (2019). *Behind the screen*. Yale University Press.

Rozado, David (2023, 2 de febrero). The unequal treatment of demographic groups by ChatGPT/OpenAI content moderation system [Substack newsletter]. *Rozado's Visual Analytics*. https://davidrozado.substack.com/p/openaicms

Salib, Peter (2022). Big data affirmative action (SSRN Scholarly Paper 4024623). *SSRN 4024623*. https://papers.ssrn.com/abstract=4024623

SimilarWeb (2024). *Chatgpt.com traffic analytics, ranking & audience [August 2024]*. Similarweb. https://www.similarweb.com/website/chatgpt.com/

Sorbán, Kinga (2021). Ethical and legal implications of using AI-powered recommendation systems in streaming services. *Információs Társadalom*, *21*, 63. https://doi.org/10.22503/inftars.XXI.2021.2.5

Suanzes, Pablo (2017, 27 de junio). *La UE impone a Google una multa récord de 2.420 millones por abuso de posición dominante*. ELMUNDO. https://www.elmundo.es/economia/macroeconomia/2017/06/27/595229ff268e3e5a578b458b.html

Tabarés Gutiérrez, Raúl (2016). El surgimiento de HTML5; un nuevo paradigma en los estándares Web. *Teknokultura. Revista de Cultura Digital y Movimientos Sociales*, *13* (1), 169-192. https://doi.org/10.5209/rev_TK.2016.v13.n1.52152

Tabarés Gutiérrez, Raúl (2024). Plataformización, automatización y aceleración en los medios sociales. *Daimon Revista Internacional de Filosofia*, *93*, 137-152. https://doi.org/10.6018/daimon.612051

Tabarés Gutiérrez, Raúl (2025, en prensa). Interfaces conversacionales, tecnolenguajes y tecnodesigualdades. *Recerca. Revista de Pensament i Anàlisi.*

Tabarés, Raúl (2021). HTML5 and the evolution of HTML; tracing the origins of digital platforms. *Technology in Society*, *65*, 101529. https://doi.org/10.1016/j.techsoc.2021.101529

Wachter, Sandra (2024). Limitations and loopholes in the EU AI Act and AI Liability Directives: What this means for the European Union, the United States, and beyond. *Yale Journal of Law and Technology*, *26*(3). https://ora.ox.ac.uk/objects/uuid:0525099f-88c6-4690-abfa-741a8c057e00

Wachter, Sandra, Mittelstadt, Brent y Russell, Chris (2024). Do large language models have a legal duty to tell the truth? (SSRN Scholarly Paper 4771884). *SSRN 4771884*. https://doi.org/10.2139/ssrn.4771884

Watkins, Ali (2024, 6 de diciembre). Why wouldn't ChatGPT say this dead professor's name? *The New York Times*. https://www.nytimes.com/2024/12/06/us/david-mayer-chatgpt-openai.html

Yang, Zeyi (2022). *La censura China borra la plaza de Tiananmén en su IA de creación de imágenes*. MIT Technology Review. http://www.technologyreview.es/s/14577/la-censura-china-borra-la-plaza-de-tiananmen-en-su-ia-de-creacion-de-imagenes

York, Jillian C (2022). *Silicon values: The future of free speech under surveillance capitalism*. Verso Books.