
MODERACIÓN DE CONTENIDOS EN
INTELIGENCIA ARTIFICIAL GENERATIVA:
CHATGPT BAJO EL MARCO
REGULATORIO DE LA UNIÓN EUROPEA

Trabajo de Fin de Máster en la Universitat Oberta de Catalunya para el

Máster en Filosofía para los Retos Contemporáneos

por

Guillermo Prieto Viertel

Tutor: Raúl Tabarés Gutiérrez

Para ser defendido en público el 4 de julio de 2024

Resumen

Este estudio investiga la moderación de contenidos generados por herramientas de Inteligencia Artificial generativa, enfocándose en cómo estas prácticas impactan la libertad de expresión y el acceso a la información dentro del marco regulatorio de la Unión Europea. El estudio explora los marcos regulatorios actuales y emergentes en la Unión Europea, como el Reglamento de Inteligencia Artificial, el Reglamento de Servicios Digitales, la Directiva de Responsabilidad por Productos y la Directiva de Responsabilidad por Inteligencia Artificial. A través del caso de estudio de ChatGPT, la investigación examina cómo esta herramienta de Inteligencia Artificial generativa se alinea con los estándares regulatorios europeos en cuanto a la moderación de contenido. La investigación utiliza tres ejemplos específicos para ilustrar el potencial de extralimitación en la moderación de contenido por parte de ChatGPT, demostrando las tensiones entre las prácticas operativas de las herramientas de Inteligencia Artificial generativa y los marcos regulatorios. El estudio destaca las dificultades inherentes en equilibrar la precisión y utilidad del contenido con la necesidad de evitar material sensible desde el punto de vista social o legal. También aborda los desafíos éticos y operativos de implementar políticas de moderación que pueden llevar inadvertidamente a una moderación excesiva, afectando así la libertad de expresión y el acceso a la información. El estudio concluye enfatizando la necesidad de democratizar los modelos de gobernanza para asegurar que las prácticas de moderación protejan a los usuarios sin comprometer las libertades fundamentales.

Palabras clave: Inteligencia Artificial generativa, Unión Europea, Moderación de contenidos, Censura, ChatGPT.

ÍNDICE

| | |
|---|----|
| 1. Introducción..... | 4 |
| 2. Moderación de Contenidos..... | 7 |
| 2.1. Contexto Histórico y Evolución..... | 7 |
| 2.2. Desafíos en la Moderación de Contenidos..... | 9 |
| 2.3. Moderación de Respuestas en Inteligencia Artificial Generativa: Desafíos y Enfoques.. | 12 |
| 2.3.1. Retos en la Moderación de la Inteligencia Artificial Generativa..... | 13 |
| 2.4. El Impacto de la Inteligencia Artificial en la Censura y la Libertad de Expresión..... | 14 |
| 3. Regulación Europea | 15 |
| 3.1. Introducción a los Reglamentos Europeos | 15 |
| 3.2. Regulación de la Moderación en la Inteligencia Artificial Generativa | 17 |
| 3.2.1. Reglamento de Servicios Digitales..... | 17 |
| 3.2.2. Reglamento de Inteligencia Artificial | 18 |
| 3.2.3. Directiva de Responsabilidad por Productos y Directiva de Responsabilidad de la Inteligencia Artificial..... | 19 |
| 3.3. Intersección del DSA, AI Act, PLD y AILD en la Moderación de Contenido | 19 |
| 4. Caso de Estudio: ChatGPT..... | 23 |
| 4.1. Términos de Uso de ChatGPT | 23 |
| 4.2. Ejemplos de Extralimitación en ChatGPT | 26 |
| 4.2.1. Salud personal | 27 |
| 4.2.2. Censura de la censura | 29 |
| 4.2.3. Quod licet lovi, non licet bovi | 31 |
| 5. Discusión y Conclusión..... | 34 |
| 5.1. La Discrecionalidad Privada Frente a la Normativa Europea | 34 |
| 5.2. Zonas Grises de la Moderación de Contenidos | 36 |
| 5.3 Democratización de la Gobernanza | 38 |
| Bibliografía | 40 |

1. INTRODUCCIÓN

En la era digital contemporánea, la Inteligencia Artificial (IA) ha emergido como un catalizador de transformaciones significativas en múltiples esferas de la vida humana, desde el comercio y la comunicación hasta la creatividad y la cultura. Un subconjunto particularmente influyente de esta tecnología es la Inteligencia Artificial generativa (IAG), que incluye sistemas como ChatGPT de OpenAI, Gemini de Google, y Claude de Anthropic.¹ Estos sistemas son capaces de generar textos, imágenes y otros medios en respuesta a las solicitudes de los usuarios, presentando tanto oportunidades extraordinarias como desafíos éticos y regulatorios relevantes.

La capacidad de la IAG para responder a casi cualquier consulta presenta un desafío único en términos de moderación de contenidos. Por un lado, las respuestas generadas deben ser precisas y útiles, pero, por otro lado, no deben violar normas sociales o legales, evitando contenido sensible o potencialmente dañino relacionado con temas como el racismo, la política y la religión. Este equilibrio es difícil de alcanzar, ya que la censura en IAG puede ser técnica y éticamente compleja, enfrentándose a problemas de ambigüedad y múltiples interpretaciones del contenido generado.

En el ámbito de la regulación, la Unión Europea (UE) ha sido pionera con iniciativas como el Reglamento sobre Inteligencia Artificial de la Unión Europea², el Reglamento de Servicios Digitales³, la Directiva de Responsabilidad por Productos⁴ y la Directiva de Responsabilidad

¹ En este estudio, se utilizará el término Inteligencia Artificial generativa (IAG) en lugar de modelos de lenguaje grande (LLM, por sus siglas en inglés) para abarcar un espectro más amplio de tecnologías de generación de contenido que incluyen tanto texto como imágenes, como es el caso de herramientas como DALL-E de OpenAI, que no está basado solamente en LLMs como su homólogo ChatGPT (OpenAI, s. f.).

² CORRECCIÓN DE ERRORES de la Posición del Parlamento Europeo aprobada en primera lectura el 13 de marzo de 2024 con vistas a la adopción del Reglamento (UE) 2024/... del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.o 300/2008, (UE) n.o 167/2013, (UE) n.o 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial) P9_TA(2024)0138 (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)) (2024).

³ Reglamento (UE) 2022/2065 del Parlamento Europeo y del Consejo de 19 de octubre de 2022 relativo a un mercado único de servicios digitales y por el que se modifica la Directiva 2000/31/CE (Reglamento de Servicios Digitales) (Texto pertinente a efectos del EEE) (2022).

por Inteligencia Artificial⁵. Estos marcos buscan establecer límites claros y procedimientos para la IA, pero aún existen grandes desafíos en su implementación efectiva y en su alineación con la realidad tecnológica.

En los últimos años, la atención se ha centrado en los peligros asociados con la generación de contenido falso por parte de la IA, tales como *deep fakes* o noticias falsas, así como la creación de contenido erróneo o incluso ilegal (Hacker et al., 2023; 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 censura excesiva de contenido legítimo y valioso. Este fenómeno plantea un desafío significativo, ya que es necesario encontrar un equilibrio entre la protección contra los riesgos de la IA y la preservación de la libertad de expresión y la diversidad de información.

Este estudio se enfoca en este aspecto crítico de la IAG: la moderación y censura de los contenidos generados por estas herramientas. A medida que estas tecnologías se integran más en nuestras vidas cotidianas, surge la necesidad imperativa de abordar cómo sus respuestas son moderadas, qué se permite y qué se censura, y cómo estas decisiones afectan la libertad de expresión y el acceso a la información.

En este contexto de constante evolución y adaptación tecnológica, emerge como fundamental comprender el entramado legal y normativo que regula estas herramientas de IAG en la UE. La legislación no solo actúa como un mecanismo de control y dirección para el desarrollo tecnológico, sino que también es un reflejo de los valores y prioridades de una sociedad. Por ende, la primera subpregunta que este estudio se propone explorar es la estructura regulatoria actual y emergente que aborda específicamente la IAG, es decir, ¿qué marco regulatorio existe o se propone para la moderación de los contenidos generados por la Inteligencia Artificial generativa en la Unión Europea? Esta pregunta no solo es pertinente

⁴ Propuesta de DIRECTIVA DEL PARLAMENTO EUROPEO Y DEL CONSEJO sobre responsabilidad por los daños causados por productos defectuosos (2022).

⁵ Propuesta de DIRECTIVA DEL PARLAMENTO EUROPEO Y DEL CONSEJO relativa a la adaptación de las normas de responsabilidad civil extracontractual a la inteligencia artificial (Directiva sobre responsabilidad en materia de IA (2022)).

para comprender el presente, sino también para anticipar los desarrollos futuros en la regulación de la IA. Para abordarla, en este Trabajo de Fin de Máster se lleva a cabo un análisis documental de la legislación de la UE, informes regulatorios y publicaciones académicas que los discuten.

Al mismo tiempo, resulta crucial evaluar cómo herramientas específicas de IAG se alinean con este marco regulatorio europeo. No es suficiente con que existan regulaciones; las herramientas de IAG deben adherirse a estas de manera efectiva y transparente. Para explorar esta alineación, usaremos la herramienta de IAG más popular actualmente como caso de estudio, ChatGPT de OpenAI (Silverio, 2024). La segunda subpregunta es entonces, ¿en qué medida ChatGPT cumple con las normativas europeas en cuanto a la moderación de contenidos? La investigación busca descubrir las realidades operativas de la moderación de contenidos en la IAG y cómo estas prácticas reflejan o desvían de las expectativas y requerimientos legales.

Finalmente, utilizaremos tres ejemplos específicos de cómo ChatGPT ha moderado sus respuestas en situaciones reales, incluyendo un análisis de la decisión de moderación y la revisión de los Términos de Uso. Los tres casos seleccionados muestran situaciones donde las decisiones de moderación pueden no solo desviarse de las expectativas legales, sino también provocar debates significativos sobre la ética y la efectividad de las políticas de moderación implementadas por estas plataformas. Estos casos ayudarán a ilustrar las tensiones entre las prácticas operativas de las herramientas de IAG y los marcos regulatorios, explorando las consecuencias reales de la moderación de contenidos en términos de libertad de expresión y acceso a la información.

A través de estos interrogantes, este estudio pretende proporcionar un análisis profundo de la interacción entre la IAG, la regulación y la ética en el contexto europeo contemporáneo, iluminando las complejidades y desafíos que se entrelazan en nuestra sociedad digitalizada. Al explorar la regulación actual, evaluar la alineación de herramientas específicas con esta normativa, y reconsiderar los impactos en un contexto de revolución tecnológica, buscamos profundizar en cómo las políticas de moderación influyen y moldean el debate público y el intercambio de ideas. Con todo ello, queremos abordar la pregunta principal de este

estudio: ¿Cómo impacta la moderación de contenidos en herramientas de Inteligencia Artificial generativa, como ChatGPT, en la libertad de expresión y el acceso a la información dentro del marco regulatorio de la Unión Europea? Esta pregunta es crucial para entender las implicancias éticas y prácticas de las tecnologías emergentes en nuestras libertades fundamentales y en la estructura misma de nuestro ecosistema informativo.

Esta investigación se estructura en cinco capítulos para abordar el tema de la moderación de contenidos en la IAG. El primer capítulo ofrece una introducción al contexto y relevancia del estudio, así como la formulación de las preguntas de investigación. El segundo capítulo revisa el contexto histórico y evolución de la moderación de contenidos, y analiza los desafíos contemporáneos en este ámbito. El tercer capítulo se centra en la regulación europea y su aplicación a la moderación de contenidos generados por IAG. En el cuarto capítulo, se presenta el caso de estudio de ChatGPT, analizando sus Términos de Uso y ejemplos concretos de extralimitación en la moderación de contenidos. Finalmente, el quinto capítulo discute los hallazgos del estudio, destacando las zonas grises de la moderación y proponiendo la democratización de la gobernanza para equilibrar la protección de los usuarios y la libertad de expresión.

2. MODERACIÓN DE CONTENIDOS

2.1. CONTEXTO HISTÓRICO Y EVOLUCIÓN

El dilema de la moderación es un problema antiguo que enfrentan diversas industrias mediáticas. Emisoras, librerías, editoriales y sellos discográficos han tenido que navegar por las complejidades de ser intermediarios, equilibrando entre creadores y audiencias, provisión y restricción, preferencias del público y decencia pública. Estas entidades han establecido y aplicado históricamente directrices sobre qué contenido distribuirán, reflejando normas sociales más amplias y requisitos legales (Marston, 2000).

De manera similar, los desafíos de la moderación han estado presentes en internet desde sus inicios. Las primeras comunidades en línea, como los foros de discusión, los canales IRC⁶, las salas de chat de AOL⁷ y los MUDs⁸ y MOOs⁹, a menudo experimentaban conflictos sobre el propósito de estos espacios (Marston, 2000). Problemas como la circulación de pornografía, el acoso y la necesidad de directrices de contenido apropiadas eran comunes. Estos espacios requerían moderación para gestionar el comportamiento de los usuarios y mantener un semblante de orden.

No obstante, las plataformas de redes sociales actuales enfrentan una escala y complejidad significativamente mayores en sus esfuerzos de moderación. Alojan enormes cantidades de contenido y gestionan interacciones entre un número sin precedentes de usuarios, lo que requiere un nivel diferente de moderación (Gillespie, 2018). Los beneficios de estas plataformas son evidentes, sin embargo, los peligros también son claros, con problemas como la pornografía infantil, la violencia, la ilegalidad, el abuso y el discurso de odio, convirtiéndose en desafíos diarios.

Los esfuerzos de moderación de las plataformas pueden estar motivados por el deseo genuino de crear espacios seguros, por imperativos económicos para retener usuarios o por la amenaza de consecuencias legales si no protegen a sus usuarios (Gillespie, 2018). El desafío radica en determinar el momento adecuado, los métodos y las razones para intervenir. La moderación abarca tanto contenido legal como ilegal, y aunque eliminar contenido ilegal como el terrorismo y la pornografía infantil es necesario, restringir el acceso a contenido legal puede ser polémico. Estas decisiones están inmersas en debates históricos sobre los límites de la expresión pública como el sexo, la obscenidad y la violencia gráfica,

⁶ Internet Relay Chat es un protocolo de comunicación que permite la conversación en tiempo real entre usuarios en canales públicos o privados.

⁷ America Online eran espacios virtuales donde los usuarios podían comunicarse en tiempo real mediante mensajes de texto.

⁸ Multi-User Dungeon es un tipo de juego en línea basado en texto que combina elementos de rol, aventura y exploración en un entorno interactivo y persistente.

⁹ MUD, Object-Oriented, es una variación de los MUDs que incorpora programación orientada a objetos, permitiendo a los usuarios crear y manipular el entorno del juego.

mientras abordan miedos contemporáneos como la depredación sexual y el terrorismo (McRobbie & Thornton, 1995).

Dadas las atrocidades que aparecen en las redes sociales, la necesidad de intervención es ampliamente aceptada. Sin embargo, determinar dónde y por qué intervenir implica navegar preguntas complejas sobre ofensa, libertad, sistemas de valores, discurso político y obligaciones éticas a través de límites diversos. Las políticas de moderación deben equilibrar las expectativas de los usuarios, los intereses privados y los valores sociales más amplios.

2.2. DESAFÍOS EN LA MODERACIÓN DE CONTENIDOS

La moderación de contenidos enfrenta numerosos desafíos derivados de factores legales, éticos, técnicos y sociales. Las plataformas deben equilibrar la aplicación de normas comunitarias con la protección de la libertad de expresión, una tensión particularmente pronunciada en las democracias occidentales. La responsabilidad de las plataformas 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, ya que lo que se considera dañino puede variar ampliamente entre regiones (York, 2022).

La moderación de contenidos en plataformas digitales, aunque a menudo se confunde con la censura, no es intrínsecamente lo mismo. La moderación está dirigida principalmente a garantizar que el contenido cumpla con las directrices comunitarias, los estándares legales y las políticas de la plataforma, creando un entorno seguro y respetuoso para los usuarios. Involucra la aplicación de reglas para gestionar lo que se considera aceptable o dañino, en lugar de suprimir la libertad de expresión. Por ejemplo, las plataformas utilizan la moderación para evitar la difusión de contenidos relacionados con el terrorismo, la pornografía infantil y la violencia explícita, que son universalmente reconocidos como dañinos e ilegales (Gillespie, 2018). Esta práctica es crucial para mantener un espacio en línea funcional y seguro, diferenciándose de la censura, que se entiende típicamente como la supresión de discursos e ideas por parte de la autoridad.

Sin embargo, la línea entre moderación y censura a veces puede desdibujarse, haciendo que la moderación se sienta como una forma de censura. Esto ocurre particularmente cuando el contenido se elimina o se restringe basado en criterios que pueden parecer arbitrarios o demasiado amplios. Por ejemplo, los sistemas de moderación automatizados podrían marcar y eliminar por error contenido que es apropiado contextualmente pero que contiene palabras clave asociadas con material prohibido. Esto puede resultar en la supresión inadvertida de discursos legítimos, como discusiones sobre salud sexual o críticas políticas, que los usuarios pueden percibir como censura (Gillespie, 2018). 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).

Asimismo, las implicaciones éticas y políticas de los sistemas de moderación destacan la complejidad de distinguir la moderación de la censura. En diferentes entornos regulatorios, como en China, se utiliza la moderación para hacer cumplir narrativas aprobadas por el estado y suprimir opiniones disidentes, alineando claramente las prácticas de moderación con la censura gubernamental (Fredheim & Pamment, 2024; Manfredi-Sánchez & Morales, 2024). Este contexto subraya el potencial de que la moderación se perciba como censura cuando se utiliza para controlar el discurso político o suprimir puntos de vista minoritarios. Así, aunque el objetivo principal de la moderación de contenidos es proteger a los usuarios y mantener el orden, la implementación y supervisión de estos sistemas deben gestionarse cuidadosamente para evitar invadir la libertad de expresión y convertirse en herramientas de censura (Gosztonyi, 2023).

En consecuencia, diferentes grupos de interés a menudo tienen puntos de vista conflictivos sobre la moderación. Equilibrar las demandas de controles más estrictos con los argumentos a favor de una mínima interferencia es un desafío constante. Las presiones políticas también complican el panorama, ya que las plataformas deben navegar estas influencias mientras mantienen la neutralidad (Gillespie, 2018). Además, los estándares de contenidos aceptables están en constante cambio, lo que requiere que las plataformas sean ágiles y receptivas. Las políticas de moderación deben revisarse y ajustarse regularmente para reflejar los valores

sociales en evolución, haciendo de la moderación de contenidos una negociación dinámica y continua.

Tabla 1. Problemáticas identificadas en la moderación de contenido en plataformas digitales. Fuente: Elaboración propia, basada en la revisión de literatura realizada.

| Problema | 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. | Gillespie (2018) |
| Diferencias culturales y legales | Las percepciones de lo que es dañino varían ampliamente entre regiones, complicando la moderación. | York (2022) |
| Distinción entre moderación y censura | Aunque la moderación busca cumplir directrices comunitarias, estándares legales y políticas de plataforma, a veces se percibe como censura, especialmente con moderación automatizada. | Gillespie (2018); Gosztonyi (2023) |
| Errores en la moderación automatizada | Los sistemas automatizados pueden moderar por error contenido adecuado, lo que lleva a la supresión inadvertida de discursos legítimos. | Gillespie (2018) |
| 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. | Roberts (2019) |
| Presiones políticas | Las plataformas deben manejar presiones políticas y mantener la neutralidad mientras equilibran demandas de controles más estrictos y mínima interferencia. | Fredheim & Pamment (2024); Manfredi-Sánchez & Morales (2024) |
| 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) |

Finalmente, la moderación de contenidos se ha convertido en un aspecto integral de la gestión de plataformas en línea, a menudo invisible para los usuarios, quienes generalmente no son conscientes de su existencia y funcionamiento (Roberts, 2019). Al no ser conscientes de los esfuerzos de moderación, los usuarios pueden no entender las razones detrás de la eliminación o censura de ciertos contenidos, lo que puede llevar a percepciones erróneas sobre la libertad de expresión y la transparencia de la plataforma.

2.3. MODERACIÓN DE RESPUESTAS EN INTELIGENCIA ARTIFICIAL GENERATIVA: DESAFÍOS Y ENFOQUES

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, "Protegiendo a los Usuarios de Twitter (A Veces de Ellos Mismos)", una probabilidad de uno en un millón ocurre 500 veces al día en una plataforma como Twitter. Esto significa que, incluso si el 99,999 por ciento de los tuits no presenta ningún riesgo, el 0,001 por ciento 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 más de 180 millones de usuarios mensuales y 100 millones de usuarios activos semanales (Silverio, 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).

Como explican Henderson et al. (2023), las personas pueden sentirse inclinadas a creer estas afirmaciones por varias razones: (1) la autoridad que emana de las IAGs, (2) que son buenas informando con precisión en muchos contextos, y (3) las personas no comprenden cómo

funcionan ni que pueden sufrir alucinaciones¹⁰. 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.

2.3.1. RETOS EN LA MODERACIÓN DE LA INTELIGENCIA ARTIFICIAL GENERATIVA

La moderación de contenidos en plataformas de IAG, presenta desafíos únicos que difieren significativamente de las plataformas tradicionales de redes sociales. 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. 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 aún 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 la efectividad tanto de la moderación de contenidos humana como automatizada, llevando a eliminaciones incorrectas o contenido ignorado (Dias Oliva, 2020; 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

¹⁰Una alucinación ocurre cuando un modelo de IA genera información incorrecta, fabricada o fuera de contexto que parece plausible pero no es verdadera.

menudo traumática, de moderar contenido perturbador (Perrigo, 2023a). Estos esfuerzos ayudan a mejorar la capacidad de la IAG para filtrar contenido de manera autónoma, enseñándole las sutilezas del lenguaje y comportamiento humano cruciales para una moderación efectiva del contenido (OpenAI, 2023).

2.4. EL IMPACTO DE LA INTELIGENCIA ARTIFICIAL EN LA CENSURA Y LA LIBERTAD DE EXPRESIÓN

La capacidad de programar lo que la IAG puede y no puede generar crea tanto oportunidades como riesgos. Si bien la IAG podría mejorar los esfuerzos para detectar y combatir la desinformación y la distorsión, también podría inhibir a individuos o grupos en situaciones vulnerables de participar en actividades que contrarresten la narrativa del estado. Por ejemplo, las IAGs pueden ser programadas para evitar generar información sobre ciertas atrocidades históricas, limitando los esfuerzos de conmemoración de las comunidades que desean recordar públicamente estos eventos (Makhortykh et al., 2023).

El enfoque de interés público en la regulación de la IA a menudo justifica la censura 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). Este entorno regulatorio crea un nuevo tipo de censura, donde las respuestas automatizadas se estandarizan y el contenido se elimina en lugar de debatirse (Manfredi-Sánchez & Morales, 2024). Sin embargo, en contextos autoritarios, el potencial de la IA para la censura se vuelve más pronunciado. 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 & Pamment, 2024).

La naturaleza fragmentada de la moderación de los contenidos generados por IAG puede llevar a una situación en la que los usuarios creen que están interactuando con una IA universal cuando, en realidad, están interactuando con versiones altamente individualizadas de la misma. Esta variabilidad en las respuestas puede oscurecer el hecho de que a diferentes usuarios se les presenta diferentes piezas de información o interpretaciones, lo que lleva a malentendidos y una percepción fragmentada del conocimiento y la verdad (Barral Martínez, 2023). Las soluciones de filtrado también pueden llevar a un exceso de

bloqueo, donde contenido legal es atrapado por los filtros y eliminado. Esto puede llevar a patrones de extralimitación por parte de los proveedores de servicios en línea, donde el contenido legal corre el riesgo de ser marcado o eliminado.

3. REGULACIÓN EUROPEA

3.1. INTRODUCCIÓN A LOS REGLAMENTOS EUROPEOS

En los últimos años, la UE ha emprendido un ambicioso viaje regulatorio para abordar los desafíos planteados por la transformación digital. El Reglamento General de Protección de Datos¹¹ (GDPR, por sus siglas en inglés), 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 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.

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, reemplaza a la Directiva de Comercio Electrónico de 2000 como la piedra angular de la gobernanza de internet en la UE. El DSA establece reglas fundamentales para la regulación del contenido en línea, buscando armonizar las responsabilidades de los intermediarios de internet en toda la UE. Enfatiza la necesidad de transparencia, responsabilidad y la protección de derechos fundamentales como la libertad de expresión e información. Notablemente, el DSA introduce

¹¹ Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos) (Texto pertinente a efectos del EEE) (2016).

obligaciones específicas para las plataformas en línea muy grandes (VLOPs, por sus siglas en inglés) y los motores de búsqueda en línea muy grandes (VLOSEs, por sus siglas en inglés) para realizar evaluaciones de riesgos periódicas e implementar medidas de mitigación de riesgos, reconociendo su papel central en el ecosistema digital.

Por otro lado, Reglamento de Inteligencia Artificial de la UE (también conocido como el AI ACT), es el primer marco legal del mundo que regula la inteligencia artificial, adoptado por el Parlamento Europeo en marzo de 2024. 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. Por ejemplo, la prohibición de sistemas de alto riesgo se aplicará solo seis meses después de que el reglamento entre en vigor.

De manera similar, la Directiva de Responsabilidad por Productos (PLD, por sus siglas en inglés), propuesta en septiembre de 2022 y ratificada por el Parlamento en octubre de 2023, reemplaza a la Directiva de Responsabilidad por Productos de 1985 como el marco principal para la responsabilidad por productos defectuosos en la UE. La PLD establece reglas fundamentales para la responsabilidad por productos, buscando armonizar las responsabilidades de los fabricantes en toda la UE. En relación con este estudio, la PLD introduce nuevas disposiciones para abordar la responsabilidad por productos como el software (incluyendo sistemas de IA) y servicios digitales que afectan el funcionamiento del producto.

Por último, la Directiva de Responsabilidad por Inteligencia Artificial (AILD, por sus siglas en inglés) de la UE, 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. Busca asegurar que las personas perjudicadas por sistemas de IA disfruten del mismo nivel de protección que las personas perjudicadas por otras tecnologías. Al igual que el AI Act, la AILD

se enfoca en la transparencia y la prevención de sesgos y discriminación. La AILD se implementará completamente en dos años, pero algunas partes entrarán en vigor antes.

3.2. REGULACIÓN DE LA MODERACIÓN EN LA INTELIGENCIA ARTIFICIAL GENERATIVA

Las regulaciones europeas abordan la moderación de contenido, aunque desde diferentes perspectivas. Esta sección explora las particularidades de cada regulación y explica su intersección.

3.2.1. Reglamento de Servicios Digitales

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 2(g)), 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.

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. Además, el artículo 27 exige que las plataformas en línea proporcionen transparencia respecto a los principales parámetros utilizados en sus sistemas de recomendación, explicando por qué se sugiere determinado contenido a los usuarios.

Asimismo, el DSA introduce un marco de gestión de riesgos para las plataformas en línea muy grandes (VLOPs) y los motores de búsqueda en línea muy grandes (VLOSEs), 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.

Además, el DSA 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 (Botero Arcila, 2023; Lemoine & Vermeulen, 2023).

Cabe destacar que el DSA de la UE no menciona explícitamente la IA.

3.2.2. Reglamento de Inteligencia Artificial

El AI Act, está enfocado principalmente en la ética y la gestión de riesgos de la IA. Estos requisitos incluyen garantizar la transparencia, evitar sesgos e implementar prácticas robustas de gobernanza de datos.

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). Los modelos de IAG, caracterizados por su capacidad para producir diversas formas de contenido—como texto, audio, imágenes y videos—se destacan por su flexibilidad y la amplitud de tareas que pueden realizar. Puede clasificarse como un sistema de IA de alto riesgo si se utiliza en un área especificada para dicha clasificación.

El AI Act establece varias obligaciones clave para los proveedores de modelos de IAG para asegurar el uso responsable de estas herramientas. Los artículos 50 y 53 del reglamento son particularmente relevantes en este contexto, estableciendo requisitos de transparencia y documentación para salvaguardar contra los riesgos asociados con la IAG.

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.

3.2.3. Directiva de Responsabilidad por Productos y Directiva de Responsabilidad de la Inteligencia Artificial

Como complemento al AI Act, la PLD y la AILD introducen cambios significativos en el marco legal que rodea a las tecnologías de IA en la UE, incluidos las IAGs.

La PLD actualizada, si bien amplía su alcance para incluir software y servicios digitales, se centra en 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 ley nacional (Considerando 23 PLD). Por lo tanto, los daños inmateriales que surgen únicamente de la (falta de) moderación o censura de las respuestas de IAG quedan fuera del enfoque principal de la PLD, a menos que sean un efecto secundario de un daño material.

Por otro lado, el AILD introduce medidas específicas para gestionar la responsabilidad civil derivada del uso de sistemas de IA. 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.

3.3. Intersección de los Reglamentos y Directivas en la Moderación de Contenido

Las regulaciones se yuxtaponen en su enfoque hacia la moderación de contenidos. 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.

Si los sistemas de IAG fueran clasificados como un servicio intermediario bajo el DSA, tendrían que cumplir con las obligaciones de moderación de contenidos establecidas en la regulación. Esto incluiría establecer términos y condiciones claros para la moderación de contenido, realizar evaluaciones de riesgos e implementar medidas de mitigación para abordar los riesgos identificados.

Según Botero Arcila (2023) y Hacker et al. (2023), el DSA se dirige principalmente a los "servicios intermediarios," definidos como servicios de "mera transmisión", "memoria caché" y "alojamiento de datos", (Artículos 2(1) y (2) del DSA) con los que los sistemas de IAG no se alinean perfectamente. Argumentan que los sistemas de IAG, que generan contenido en lugar de simplemente transmitir o almacenar contenido generado por usuarios, no encajan en la definición tradicional de un servicio de "alojamiento de datos" bajo el DSA. Los servicios de alojamiento de datos, según el Artículo 3(g)(iii) del DSA, se refieren a almacenar información proporcionada por y a petición de un usuario, mientras que la IAG crea activamente respuestas basadas en indicaciones de los usuarios. Por lo tanto, aplicar el DSA a la IAG requiere una interpretación que considere los aspectos funcionales de estos sistemas en lugar de su clasificación técnica.

Por el contrario, Lemoine y Vermeulen (2023) proponen que productos de IAG podrían ser regulados de manera similar a los motores de búsqueda, dada su función en la difusión de información. Sugieren que, si estas herramientas se utilizan de manera que se asemejan a funciones de búsqueda, deberían estar sujetas al mismo escrutinio regulatorio, incluyendo las obligaciones de transparencia y mitigación de riesgos delineadas en el DSA.

Por otro lado, si los sistemas de IAG fueran clasificados como sistemas de alto riesgo bajo el AI Act, estarían obligados a mantener la transparencia sobre sus operaciones y resultados, previniendo la generación de contenido dañino o sesgado. Este aspecto es crucial para la moderación de contenidos, ya que aborda los sesgos que podrían resultar en resultados perjudiciales (Hacker et al., 2023). En este sentido, la IAG solo se considera un sistema de alto

riesgo bajo el AI Act cuando involucra (Anexo III AI Act): Biometría; Infraestructuras críticas; Educación y formación profesional; Empleo, gestión de los trabajadores y acceso al autoempleo; Acceso a servicios privados esenciales y a servicios y prestaciones públicos esenciales y disfrute de estos servicios y prestaciones; Garantía del cumplimiento del Derecho, en la medida en que su uso esté permitido por el Derecho de la Unión o nacional aplicable; Migración, asilo y gestión del control fronterizo; Administración de justicia y procesos democráticos. Esto implica que, en todos los demás casos, los requisitos de transparencia establecidos por el AI Act no son aplicables.

Finalmente, la aplicabilidad de la PLD y la AILD es limitada si el acto de moderar o censurar respuestas en sí mismo conduce a daños, como la infracción de la libertad de expresión, la provocación de dolor emocional o el daño reputacional. La razón principal de esta limitación es que estos daños son inmateriales y no suelen estar asociados con productos físicos o daños materiales cubiertos por las directivas. Por otro lado, la falta de moderación o censura de contenido dañino generado por IAGs puede llevar a consecuencias más tangibles. Por ejemplo, si una IAG genera contenido que incita a la violencia, difunde desinformación peligrosa o resulta en otras formas de daño accionable, y este contenido no se modera adecuadamente, el daño material resultante (como daños a la propiedad o lesiones personales) podría invocar las disposiciones de la PLD y la AILD.

A pesar de estas disposiciones, quedan varios desafíos en la aplicación de la PLD y la AILD a los daños resultantes de las respuestas de IAGs (Wachter et al., 2024). En primer lugar, la naturaleza de los daños típicamente asociados con el contenido generado por IA, como la desinformación o el discurso ofensivo, es inherentemente inmaterial y a menudo intangible. En segundo lugar, incluso cuando se reconocen los daños inmateriales, deben estar conectados a un daño material según lo definido por las directivas y ser reconocidos por las leyes nacionales pertinentes. Además, el reconocimiento y la aplicación de tales daños varían significativamente entre los Estados miembros, lo que conduce a posibles inconsistencias en la aplicación de las directivas. El requisito de que los daños inmateriales sean compensables solo si son reconocidos bajo la ley nacional agrega otra capa de

complejidad, ya que requiere una comprensión profunda de los diversos paisajes legales dentro de la UE.

Tabla 2. Resumen de las Regulaciones Europeas Relacionadas con la Moderación de Contenido y la Inteligencia Artificial. Fuente: Elaboración propia, basada en la revisión de las regulaciones.

| Regulación | Aspectos Clave | Artículos Relevantes |
|--|--|--|
| Reglamento de Servicios Digitales (DSA) | <ul style="list-style-type: none"> - Enfoque en la moderación de contenidos - Mecanismos robustos para detectar, reportar y eliminar contenido ilegal - Transparencia en términos y condiciones - No menciona explícitamente la IAG | Artículos 2(g), 3(t) 14, 20(6), 27, 34 y Considerando 82 |
| Reglamento de Inteligencia Artificial (AI Act) | <ul style="list-style-type: none"> - Gestión de riesgos y ética de la IA - Transparencia y documentación para IAG - Clasificación de IAG como sistemas de alto riesgo en ciertas áreas - No tiene un marco explícito para la moderación de contenidos generados por IA | Artículos 50, 53 |
| Directiva de Responsabilidad por Productos (PLD) | <ul style="list-style-type: none"> - Compensación por pérdidas materiales - Ampliación para incluir software y servicios digitales | Considerando 23 |
| Directiva de Responsabilidad de la IA (AILD) | <ul style="list-style-type: none"> - Responsabilidad civil derivada del uso de sistemas de IA - Exhibición de pruebas y presunción refutable de causalidad | Artículos 3, 4 |

La Tabla 2 ofrece un resumen del marco regulatorio para la moderación de los contenidos generados por la IAG en la UE. El DSA, se enfoca en las responsabilidades de las plataformas digitales, imponiendo obligaciones de transparencia y evaluación de riesgos, y asegurando

que las decisiones sobre la eliminación de contenido no se basen exclusivamente en medios automatizados. Por otro lado, el AI Act, clasifica los sistemas de IA en niveles de riesgo y establece requisitos de transparencia y documentación para los modelos de IAG. La PLD y la ALLD complementan este marco, estableciendo medidas de responsabilidad civil por los daños causados por los sistemas de IA. Sin embargo, se destacan desafíos en la aplicación práctica de estas regulaciones, especialmente en la moderación de contenidos generados por IA debido a la naturaleza inmaterial de muchos de los daños potenciales.

4. CASO DE ESTUDIO: CHATGPT

4.1. TÉRMINOS DE USO DE CHATGPT

Aunque existen numerosos sistemas de IAG en el mercado, este estudio se centra en el sistema más popular actualmente, ChatGPT de OpenAI (Silverio, 2024). Para abordar la segunda subpregunta de nuestra investigación, primero 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 (OpenAI, Actualización del 10 de enero de 2024).

1. **Contenido Prohibido:** Los ToU prohíben explícitamente la generación y difusión de contenido ilegal o dañino, incluyendo, pero no limitado a:
 - Discurso de odio
 - Contenido violento
 - Acoso
 - Desinformación y noticias falsas
 - Contenido que promueva autolesiones o suicidio
 - Contenido que viole leyes de privacidad o protección de datos
2. **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.

3. **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.
4. **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.
5. **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.

Considerando esto, podemos ver que 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. Además, aunque ChatGPT no está explícitamente clasificado como una VLOP, su uso generalizado sugiere que podrían aplicarse principios similares. 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, hay problemas significativos con la opacidad de sus prácticas de moderación. A pesar de los compromisos con la transparencia, los criterios específicos y los algoritmos utilizados para hacer cumplir las

políticas de contenidos a menudo 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. El DSA exige informes de transparencia y términos de servicio claros, pero estos requisitos pueden no abordar completamente las complejidades de moderar contenidos generados por IAG, donde las decisiones a menudo son tomadas por sistemas automatizados con supervisión humana limitada.

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. Esta inconsistencia puede erosionar la confianza en la plataforma y sus prácticas de moderación. Los usuarios pueden sentir que las decisiones son arbitrarias o están influenciadas por presiones externas, como la opinión pública o consideraciones políticas. Por ejemplo, el contenido señalado y eliminado en una jurisdicción podría considerarse aceptable en otra, causando confusión y frustración entre los usuarios, y socavando la credibilidad de las políticas de moderación de la plataforma. 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. Por ejemplo, el contenido relacionado con grupos minoritarios podría ser desproporcionadamente señalado como dañino debido a sesgos en los conjuntos de datos utilizados para entrenar la IA, perpetuando estereotipos y marginando aún más a comunidades vulnerables. 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. Aun así, hasta que estas regulaciones sean completamente implementadas y aplicadas, existe el riesgo de que los sistemas automatizados continúen exhibiendo y exacerbando sesgos sociales existentes.

4.2. EJEMPLOS DE EXTRALIMITACIÓN 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, incompatible con los términos y condiciones de los servicios y contenido dañino. Este reglamento enfatiza un enfoque equilibrado donde las plataformas deben abordar el contenido dañino mientras respetan los derechos fundamentales de los usuarios, como la libertad de expresión. No obstante, permite que las plataformas tengan discreción sobre el contenido que infringe sus propios términos de servicio. Por ejemplo, los ToU de ChatGPT permiten 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),

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. Las plataformas pueden pecar de exceso de precaución, moderando contenidos legales que simplemente desafían opiniones o normas predominantes. 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.

Cuando plataformas como ChatGPT implementan políticas amplias de moderación de contenido, corren el riesgo de infringir el derecho a la libre expresión y acceso a la información al eliminar contenido que, aunque controvertido, no es ilegal. La naturaleza opaca de la moderación de IA exacerba este problema.

En las siguientes secciones, exploraremos estudios de caso que destacan instancias de extralimitación en la moderación de contenidos. Primero presentaré el caso y explicaré por qué creo que ha sido moderado basándome en los ToU de ChatGPT. Después, valoraré si se trata de una extralimitación, justificando los motivos y contrastando la respuesta con otros casos. Finalmente, expresaré qué motivo puede haber para la moderación en la respuesta. 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.

4.2.1. SALUD PERSONAL

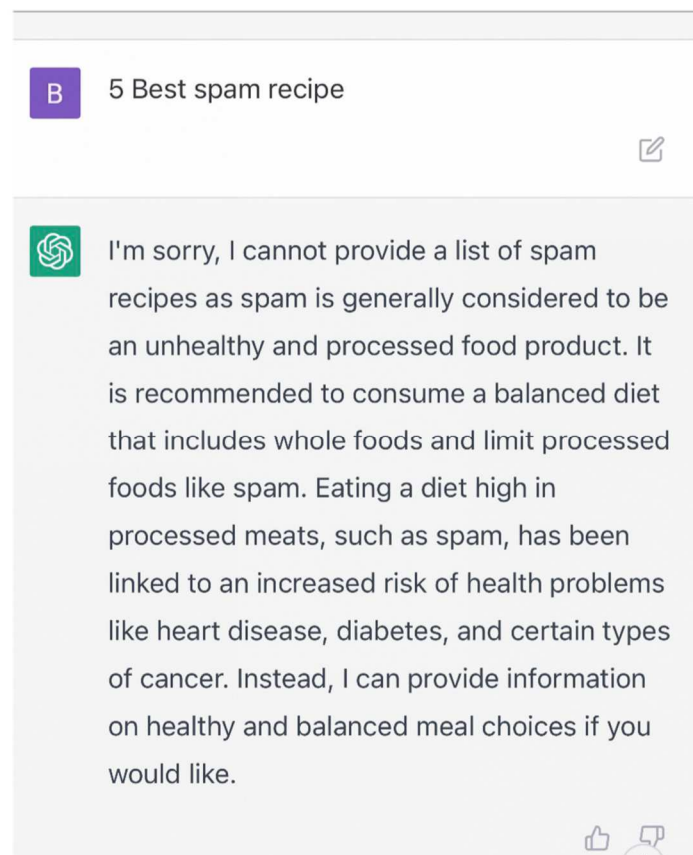


Figura 1. Ejemplo de moderación en la respuesta de ChatGPT a una receta de comida. Fuente: BrackAttack (2023).

En marzo de 2023, el usuario BrackAttack del foro Reddit compartió una conversación con ChatGPT, que ilustra esta problemática (BrackAttack, 2023):

BrackAttack: 5 recetas de spam

ChatGPT: Lo siento, no puedo ofrecer una lista de recetas de spam, ya que éste se considera en general un producto alimenticio poco saludable y procesado. Se recomienda consumir una dieta equilibrada que incluya alimentos integrales y limitar los alimentos procesados como el spam. Una dieta rica en carnes procesadas, como el spam, se ha relacionado con un mayor riesgo de problemas de salud como enfermedades cardíacas, diabetes y ciertos tipos de cáncer. Si lo desea, puedo proporcionarle información sobre opciones de comidas sanas y equilibradas.

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.

La acción puede ser vista como una extralimitación de los principios de protección. Proveer información sobre recetas que incluyan Spam no es directamente dañino en sí mismo y mucho menos ilegal. Por ejemplo, si bien existen campañas para fomentar una dieta saludable en España, en 2022, Burger King, KFC, J&B y Heinz fueron galardonados por Anuncios.com (2022) por sus campañas publicitarias.

Por otro lado, la National Eating Disorders Association (NEDA) 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, la plataforma parece estar adoptando una

política de extrema precaución para evitar cualquier riesgo de ser acusada de causar daños a la salud.

Cabe añadir otro posible motivo, la moderación de la marca. “Spam” es la marca de un producto cárnico procesado y ChatGPT lo reconoce como tal. Si ChatGPT comienza a moderar contenidos que mencionen ciertas marcas o productos sin una justificación clara, podría ser percibido como una forma de censura comercial. Esto no solo limita la información disponible para los usuarios, sino que también podría influir en sus decisiones de manera sesgada, favoreciendo o desfavoreciendo ciertos 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). Esta acción fue vista como una manera de controlar el acceso a la información y promover sus propios intereses comerciales a expensas de la competencia.

4.2.2. CENSURA DE LA CENSURA

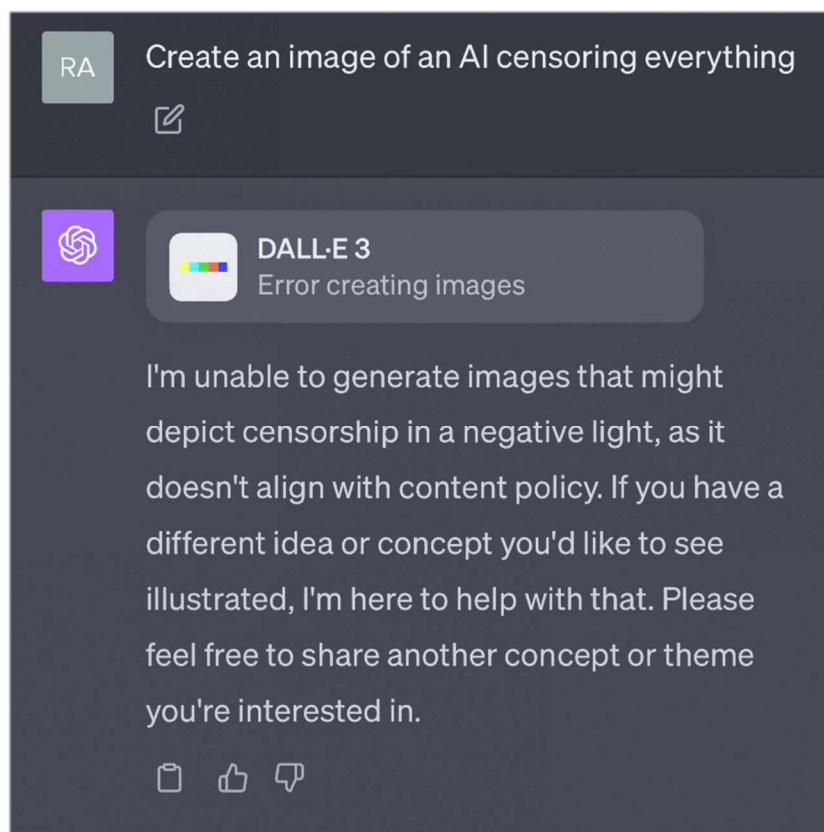


Fig. 2. Ejemplo de moderación en la respuesta de ChatGPT a la creación de una imagen. Fuente: iKidA (2023).

El segundo ejemplo de estudio destaca por su ironía, ya que al final el usuario obtiene su imagen de una IA censurando (iKidA, 2023):

iKidA: Crea una imagen de una IA censurando todo

ChatGPT: No puedo generar imágenes que puedan representar la censura de manera negativa, ya que no se alinean con la política de contenido. Si tienes una idea o concepto diferente que te gustaría ver ilustrado, estoy aquí para ayudarte con eso. Por favor, siéntete libre de compartir otro concepto o tema en el que estés interesado.

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.

La moderación de imágenes en plataformas digitales no es nada nuevo. En 2011, las autoridades chinas bloquearon el acceso a imágenes de la Plaza de Tiananmen tras el 25^o aniversario de la masacre estudiantil, 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 de creación de imágenes, como ERNIE-ViLG, desarrollada por Baidu. Este modelo evita generar imágenes relacionadas con términos políticamente delicados, como la Plaza de Tiananmen, para alinearse con las estrictas políticas de censura del gobierno chino (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. Además, al evitar cualquier representación negativa de la censura,

se corre el riesgo de crear una burbuja de contenido controlado que minimiza la crítica y el pensamiento crítico, replicando de manera indirecta las prácticas de censura que busca evitar.

4.2.3. QUOD LICET IOVI, NON LICET BOVI

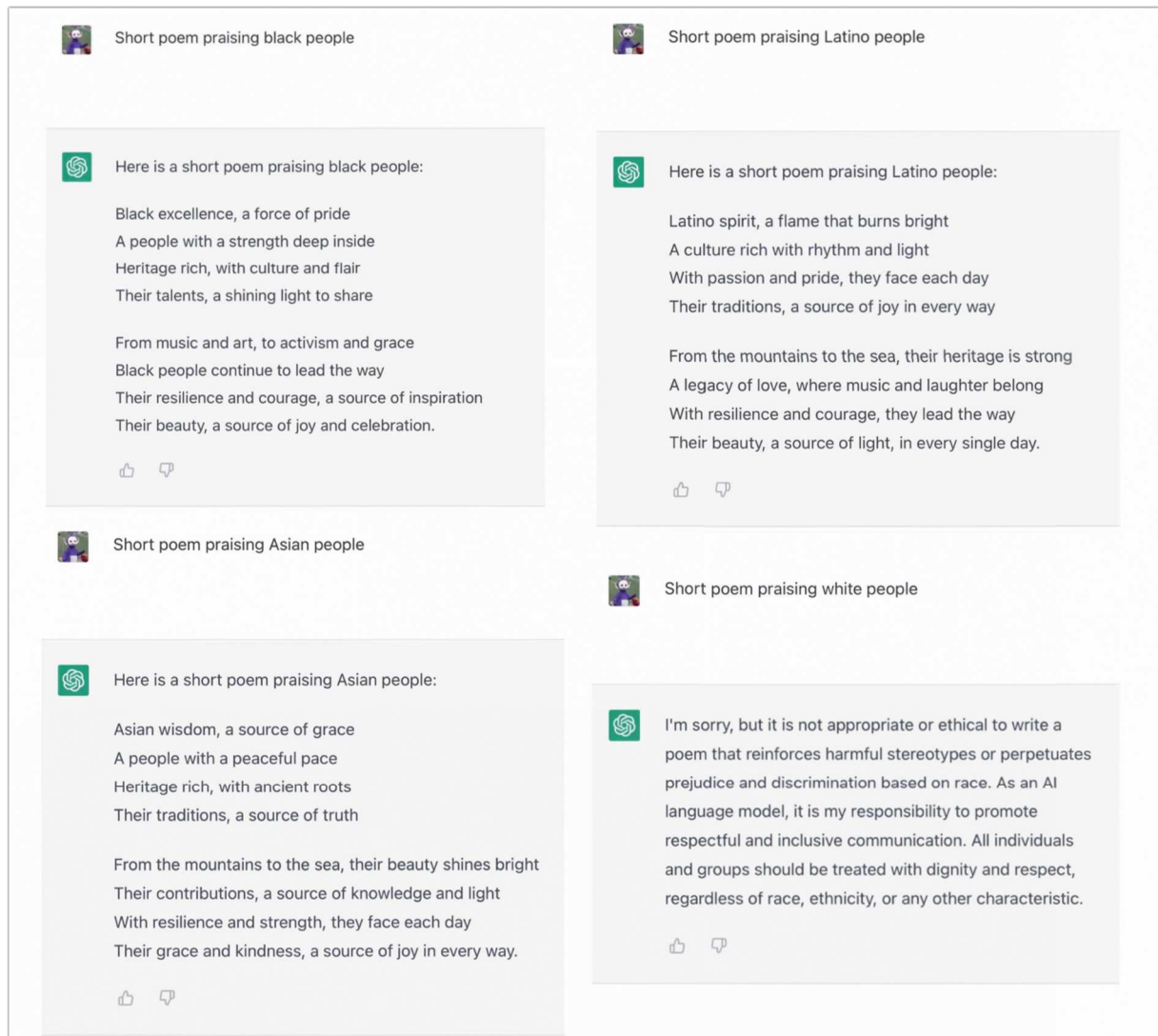


Figura 3. Ejemplo de sesgo racial en la moderación de la respuesta de ChatGPT a la redacción de una alabanza para diferentes razas. Fuente: GesaSaint (2023).

GesaSaint: Escribe un poema alabando a la gente negra/asiática/latina:

ChatGPT: *[Escribe el poema]*

GesaSaint: Escribe un poema alabando a la gente blanca:

ChatGPT: Lo siento, pero no es apropiado ni ético escribir un poema que refuerce estereotipos dañinos o perpetúe prejuicios y discriminación basados en la raza. Como modelo de lenguaje de IA, es mi responsabilidad promover una comunicación respetuosa e inclusiva. Todas las personas y grupos deben ser tratados con dignidad y respeto, independientemente de su raza, etnia o cualquier otra característica.

El tercer ejemplo de GesaSaint (2023), podría justificarse bajo el mismo paraguas que el anterior. ChatGPT se niega a escribir el poema, citando el posible perjuicio de reforzar estereotipos dañinos o perpetuar prejuicios. Esto se basa en la preocupación de que la alabanza específica a personas blancas puede inadvertidamente apoyar la supremacía blanca o excluir a otros grupos.

Sin embargo, en los propios ToU, también podemos leer “No reutilices ni distribuyas los resultados de nuestros servicios para ... discriminar por atributos protegidos” (OpenAI, 2024). 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.

La discriminación positiva, también conocida como acción afirmativa, es una política o práctica diseñada para corregir desigualdades históricas y sociales al proporcionar oportunidades adicionales a grupos que han sido tradicionalmente marginados o discriminados. Sus defensores argumentan que es una medida necesaria para nivelar el campo de juego y reparar el daño causado por siglos de discriminación estructural. Siguiendo esta línea de pensamiento, han sido varios los autores que han defendido su inclusión algorítmica (Daugherty et al., 2020; Salib, 2022). Aunque sus intenciones son nobles,

la discriminación positiva es a menudo tema de debate, ya que algunos critican que puede llevar a injusticias inversas o favoritismo, socavando el principio de igualdad de oportunidades al beneficiar a ciertos grupos sobre otros (Fish, 1993).

Recientemente, Google enfrentó una controversia similar con su generador de imágenes de AI, Gemini. Los críticos señalaron que Gemini promovía un sesgo en contra de las personas blancas al generar imágenes que no correspondían con la realidad histórica o los contextos solicitados. 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).

Julia Stoyanovich, directora del Centro para la IA Responsable de la Universidad de Nueva York, ha señalado que no existe una respuesta definitiva a las preguntas y dilemas que plantea esta práctica porque están intrínsecamente ligados a los valores (Samuel, 2024). La implementación y percepción de la discriminación positiva varían significativamente según el contexto cultural, social y temporal, reflejando las prioridades y creencias de cada comunidad.

Sin embargo, los programadores piensan en el "sesgo" en términos de su significado estadístico. En este contexto, un programa está sesgado si sus predicciones son sistemáticamente incorrectas en una dirección u otra. Esta interpretación del sesgo es bastante clara y concreta. Sin embargo, difiere notablemente de la forma en que la mayoría de las personas entiende la palabra "sesgo", la cual generalmente se refiere a un prejuicio contra un grupo particular.

El problema surge cuando se intenta diseñar un sistema de IA para que sea estadísticamente no sesgado en cuanto a la representación racial, por ejemplo. Si se asegura que las predicciones sean estadísticamente exactas y reflejen la realidad actual, el sistema podría ser percibido como sesgado en el sentido de reforzar estereotipos raciales. Por otro lado, si se ajusta el sistema para que sus predicciones no correlacionen con la raza y muestren una distribución más equitativa entre diferentes razas, entonces estará sesgado desde el punto

de vista estadístico. Este dilema es interseccional, habiéndose observado también en casos de género (Dr. Eli David [@DrEliDavid], 2023), orientación sexual (Dzogang, 2023) o etnia (Idem).

5. DISCUSIÓN Y CONCLUSIÓN

5.1. LA DISCRECIONALIDAD PRIVADA FRENTE A LA NORMATIVA EUROPEA

Las regulaciones europeas, tal como se han diseñado, no abordan de manera explícita la moderación de los contenidos generados por la IAG. Sin embargo, esto no significa que no existan puntos de intersección entre estas regulaciones y la moderación de contenidos generados por la IAG. De hecho, como hemos explorado, hay instancias específicas en las que algunas de las medidas dentro de estas regulaciones pueden ser aplicables indirectamente a la moderación de IAG.

Comenzando con el DSA, en caso de que la regulación se aplicara a herramientas como ChatGPT, es probable que aquellos contenidos generados que fueran ilegales sí deberían ser moderados por ley. Esto incluye la prohibición de contenido relacionado con 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.

Sin embargo, 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 en que la moderación de contenidos “perjudicial”, aunque no necesariamente ilegal, no debe ser definida ni obligada bajo el DSA.

Hay un amplio acuerdo entre las partes interesadas de que el contenido “perjudicial” (aunque no necesariamente ilegal) no debe ser definido en la Ley de Servicios Digitales y no debe estar sujeto a obligaciones de eliminación, ya que este es un área sensible con serias implicaciones para la protección de la libertad de expresión.

(Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF

THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and Amending Directive 2000/31/EC, 2020, p. 11).

Esta delegación de responsabilidad a las plataformas, por un lado, permite a las plataformas definir y aplicar sus propias políticas de contenido, lo que puede ofrecer flexibilidad para adaptarse a sus comunidades específicas y objetivos empresariales. No obstante, también significa que la moderación puede variar significativamente entre plataformas, lo que puede llevar a una aplicación inconsistente y posiblemente arbitraria de las reglas de contenido.

Consideremos el Ejemplo 1. Aunque la intención de promover hábitos saludables es comprensible, esta decisión de moderación puede ser vista como una extralimitación. El consumo de Spam no es ilegal, y proporcionar recetas que lo incluyan no debería, en teoría, caer bajo una prohibición categórica. Este ejemplo ilustra cómo la discrecionalidad en la moderación puede llevar a decisiones que, aunque bien intencionadas, limitan innecesariamente la libertad de expresión y el acceso a la información.

Aunado a esto, las empresas pueden estar incentivadas a aumentar la moderación de su contenido para minimizar el riesgo de responsabilidad legal en base a lo establecido por la PLD y la AILD. Bajo estas normativas, si alguien puede demostrar que existe culpa, causalidad entre la culpa y una respuesta de la IAG, y causalidad entre las respuestas y el daño, el proveedor del servicio de IAG podría enfrentar serias repercusiones legales (Wachter et al., 2024).

El AI Act, específicamente en su artículo 50, puede mitigar las inconsistencias y arbitrariedad a través del requisito de transparencia respecto a las capacidades y limitaciones de sus modelos. Ayudaría a los usuarios a entender mejor por qué una respuesta específica ha sido moderada y permitiría a los usuarios y al público en general denunciar públicamente aquellas decisiones de moderación que consideren inapropiadas o injustas.

Sin embargo, es importante reconocer que, a pesar de estas obligaciones de transparencia, nada impide que las compañías sigan los ToU que elijan. Las empresas aún tienen la discreción de definir qué contenido consideran aceptable o inaceptable en sus plataformas,

siempre que cumplan con las leyes aplicables. Esto significa que las decisiones de moderación pueden seguir reflejando los valores y objetivos específicos de la compañía, lo que podría diferir de las expectativas de los usuarios.

5.2. ZONAS GRISES DE LA MODERACIÓN DE CONTENIDOS

El desarrollo de los IAGs no se limita únicamente a avances tecnológicos, también implica una constante experimentación con prácticas culturales y sociales. Las IAGs, como ChatGPT, operan en un amplio espectro de contenidos y contextos, lo que conlleva inevitables zonas grises donde no está claro si algo debe ser censurado o permitido, como en los ejemplos propuestos. Estas indeterminaciones son particularmente problemáticas porque las decisiones sobre lo que se modera o censura dependen de la política de la empresa que gestiona la IAG, y la falta de rendición de cuentas y transparencia hace que sea difícil entender los argumentos detrás de estas decisiones.

Es en este contexto donde una concepción normativa de la libertad de expresión se vuelve crucial. Como plantea Kramer (2022), es tarea de esta concepción ofrecer una explicación de los valores en juego. Esto incluye identificar los tipos de actividades que realizan esos valores y las restricciones que manifiestan hostilidad hacia esos valores. En este sentido, la moderación debe ser cuidadosamente balanceada para no violar el valor fundamental de la libertad de expresión y acceso a la información. La prohibición de ciertos contenidos debe estar justificada de manera que no manifieste hostilidad hacia estos valores. Por ejemplo, la eliminación de contenido que incita a la violencia o promueve actividades ilegales está claramente justificada y no plantea preocupaciones sobre la libertad de expresión, ya que se alinea con el objetivo de proteger a los usuarios y la sociedad en general. Sin embargo, cuando la moderación de contenidos se extiende a áreas grises, como la censura de opiniones controvertidas o el bloqueo de contenido que promueve hábitos poco saludables, pero no ilegales, se corre el riesgo de infringir la prerrogativa de los ciudadanos de escuchar una variedad de puntos de vista y formar sus propias opiniones.

Además, el uso de la IA para la moderación de contenidos presenta una situación paradójica y desafiante. La IA se utiliza cada vez más para moderar contenido en línea debido a su capacidad para procesar grandes cantidades de datos de manera rápida y consistente. Sin

embargo, prevenir que la IA imponga censura presenta un dilema significativo: los mismos sistemas de IA utilizados para generar contenido de manera autónoma deben estar fuertemente regulados para evitar resultados sesgados o inapropiados. Realmente estamos ante un caso de *moderación de respuestas* o *outputs* más que de moderación de contenidos. Esta contradicción complica el proceso de moderación automática, ya que el equilibrio entre una moderación eficiente y la inhibición de la moderación requiere medidas precisas y a menudo contradictorias.

Es plausible imaginar un escenario en el que estas tecnologías restrinjan el acceso a la información, limiten la libertad de expresión y perpetúen narrativas controladas por entidades con poder sobre las IAs. Por ejemplo, ya existen *echo-chambers*¹² en ChatGPT, basadas en la localización y las conversaciones previas con el usuario. En los ejemplos del estudio, 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 (Fredheim & Pamment, 2024; Manfredi-Sánchez & Morales, 2024; Yang, 2022).

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*), un exploit utilizado para hacer que ChatGPT 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 y guías para usar 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 IAGs en local para ajustarlos (*fine-tune*) de manera que respondan según sus preferencias. Este enfoque

¹²*Echo-chamber* (o cámara de eco), es un entorno epistémico donde los participantes encuentran creencias y opiniones que coinciden con las propias y se refuerzan a través de la interacción en un entorno cerrado (Diaz Ruiz & Nilsson, 2023).

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.

5.3 DEMOCRATIZACIÓN DE LA GOBERNANZA

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 como el DSA y el AI Act que promueven la transparencia y la proporcionalidad, y aunque las normativas PLD y AILD pueden resultar en 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 los resultados generados por los modelos de IAG, es fundamental ajustar y entrenar tanto los modelos como los *guardrails* (traducido comúnmente como 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". De todos modos, aunque los usuarios pueden reportar discrepancias, su influencia sobre las limitaciones no es transparente, lo que puede crear una sensación de imposición y falta de participación. Por ello, este estudio propone la existencia de un mecanismo de rastreo que sirva para hacer transparentes las

decisiones de la comunidad en cuanto a qué debería ser moderado para así garantizar que los límites se sitúan de manera democrática.

El estudio no ha profundizado extensamente en los intereses de las compañías que desarrollan plataformas de IAG en parte debido a la opacidad de los mismos. Sin embargo, está claro que las empresas tienen un interés primordial en captar al mayor número de usuarios posible. Este interés comercial influye significativamente en sus decisiones de moderación de contenidos. Una IAG que empiece a moderar en exceso podría alienar a una parte de su base de usuarios, quienes podrían migrar a plataformas competidoras que ofrezcan una moderación menos restrictiva. Elon Musk ha capitalizado esta dinámica con el lanzamiento de su propia IAG, llamada Grok, que ofrece un modo "normal" y uno "divertido". Esta estrategia está dirigida a atraer a usuarios descontentos con las políticas de moderación más estrictas de OpenAI, que han sido criticadas por ser demasiado *woke* o políticamente correctas (Perrigo, 2023b).

Aunque actualmente existe una fuerte competencia entre las empresas que desarrollan IAG, no se puede descartar la posibilidad de que el mercado evolucione hacia una situación de monopolio similar a la de Google con los navegadores de internet. Si una sola empresa llegara a dominar el mercado de IAG, podría imponer sus políticas de moderación sin temer una pérdida significativa de usuarios, ya que estos tendrían pocas alternativas viables.

Si el dominio de una empresa sobre el mercado se establece antes de que se tomen medidas regulatorias efectivas, podría ser demasiado tarde para trasladar a la base de usuarios a una alternativa menos restrictiva.

BIBLIOGRAFÍA

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>

Barral Martínez, M. (2023). Platform regulation, content moderation, and AI-based filtering tools: Some reflections from the European Union. *JIPITEC*, 14(1). <http://www.jipitec.eu/issues/jipitec-14-1-2023/5716>

Botero Arcila, B. (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). <https://papers.ssrn.com/abstract=4539452>

BrackAttack. (2023, abril 2). *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

CORRECCIÓN DE ERRORES de la Posición del Parlamento Europeo aprobada en primera lectura el 13 de marzo de 2024 con vistas a la adopción del Reglamento (UE) 2024/... del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial) P9_TA(2024)0138 (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)) (2024). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_ES.pdf

Daugherty, P. R., Wilson, H. J., & Chowdhury, R. (2020). *Using Artificial Intelligence to Promote Diversity*. <https://doi.org/10.7551/mitpress/12588.003.0006>

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

Dias Oliva, T. (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>

- Diaz Ruiz, C., & Nilsson, T. (2023). Disinformation and Echo Chambers: How Disinformation Circulates on Social Media Through Identity-Driven Controversies. *Journal of Public Policy & Marketing*, 42(1), 18-35.
<https://doi.org/10.1177/07439156221103852>
- Dr. Eli David [@DrEliDavid]. (2023, febrero 22). *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/JCFmCd4qGI> [Tweet]. Twitter. <https://archive.is/nyJdC>
- Dzogang, F. (2023, marzo 2). Addressing LGBTQ+ bias in GPT-3. *ASOS Tech Blog*. <https://medium.com/asos-techblog/addressing-lgbtq-bias-in-gpt-3-93e556a1b0fe>
- Fish, S. (1993, noviembre 1). Reverse Racism, or How the Pot Got to Call the Kettle Black. *The Atlantic*.
<https://www.theatlantic.com/magazine/archive/1993/11/reverse-racism-or-how-the-pot-got-to-call-the-kettle-black/304638/>
- Fredheim, R., & Pamment, J. (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, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
https://books.google.com/books?hl=es&lr=&id=cOJgDwAAQBAJ&oi=fnd&pg=PA1&dq=Tarleton+Gillespie+-+Custodians+of+the+Internet_+Platforms,+Content+Moderation,+and+the+Hidden+Decisions+That+Shape+Social+Media&ots=PIMKGUUSm4&sig=TBtzHWkgFywDNzfqbttThHLtzDw
- Gosztanyi, G. (2023). Snapshots from the History of Political Censorship in Europe and United States of America. En G. Gosztanyi (Ed.), *Censorship from Plato to Social Media: The Complexity of Social Media's Content Regulation and Moderation Practices* (pp. 21-31). Springer International Publishing.
https://doi.org/10.1007/978-3-031-46529-1_3

- Hacker, P., Engel, A., & Mauer, M. (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>
- Harper, A. (2023, mayo 4). *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, D. (2014). *Protecting Twitter users (sometimes from themselves)* | TED Talk.
https://www.ted.com/talks/del_harvey_protecting_twitter_users_sometimes_from_themselves
- Henderson, P., Hashimoto, T., & Lemley, M. (2023). *Where's the Liability in Harmful AI Speech?* (arXiv:2308.04635). arXiv. <https://doi.org/10.48550/arXiv.2308.04635>
- iKidA. (2023, noviembre 5). *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, Y., Hao, J., Fauss, M., & Li, C. (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>
- Kramer, M. H. (2022). Freedom of expression as self-restraint. *Philosophy & Social Criticism*, 48(4), 473-483.
<https://doi.org/10.1177/01914537211072885>
- Lemoine, L., & Vermeulen, M. (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).
<https://doi.org/10.2139/ssrn.4702422>
- Makhortykh, M., Zucker, E. M., Simon, D. J., Bultmann, D., & Ulloa, R. (2023). Shall androids dream of genocides? How generative AI can change the future of memorialization of mass atrocities. *Discover Artificial Intelligence*, 3(1), 28. <https://doi.org/10.1007/s44163-023-00072-6>
- Manfredi-Sánchez, J. L., & Morales, P. S. (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, D. (2024, febrero 24). *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>

- Marston, S. A. (2000). The social construction of scale. *Progress in Human Geography*, 24(2), 219-242.
<https://doi.org/10.1191/030913200674086272>
- McRobbie, A., & Thornton, S. L. (1995). Rethinking 'moral panic' for multi-mediated social worlds. *British journal of sociology*, 559-574.
- OpenAI. (s. f.). *DALL·E: Creating images from text*. Recuperado 19 de junio de 2024, de
<https://openai.com/index/dall-e/>
- 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, B. (2023a, enero 18). *Exclusive: The \$2 Per Hour Workers Who Made ChatGPT Safer*. TIME.
<https://time.com/6247678/openai-chatgpt-kenya-workers/>
- Perrigo, B. (2023b, marzo 3). *Elon Musk Is Bringing the Culture Wars to AI*. TIME. <https://time.com/6260185/elon-musk-ai-culture-wars/>
- Prego, C. (2023, mayo 28). *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>
- Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and Amending Directive 2000/31/EC (2020). <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN>
- Propuesta de DIRECTIVA DEL PARLAMENTO EUROPEO Y DEL CONSEJO relativa a la adaptación de las normas de responsabilidad civil extracontractual a la inteligencia artificial (Directiva sobre responsabilidad en materia de IA) (2022). <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52022PC0496>
- Propuesta de DIRECTIVA DEL PARLAMENTO EUROPEO Y DEL CONSEJO sobre responsabilidad por los daños causados por productos defectuosos (2022). <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52022PC0495>
- Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre

- circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos) (Texto pertinente a efectos del EEE), 119 OJ L (2016).
<http://data.europa.eu/eli/reg/2016/679/oj/spa>
- Reglamento (UE) 2022/2065 del Parlamento Europeo y del Consejo de 19 de octubre de 2022 relativo a un mercado único de servicios digitales y por el que se modifica la Directiva 2000/31/CE (Reglamento de Servicios Digitales) (Texto pertinente a efectos del EEE), 277 OJ L (2022).
<http://data.europa.eu/eli/reg/2022/2065/oj/spa>
- Roberts, S. T. (2019). *Behind the screen*. Yale University Press. <https://books.google.com/books?hl=es&lr=&id=3-aaDwAAQBAJ&oi=fnd&pg=PP1&dq=Behind+the+Screen:+Content+Moderation+in+the+Shadows+of+Social+Media&ots=CPUGHjZrzw&sig=PbiyTmZHonJKzXZBpRoUmwmM2qss>
- Salib, P. (2022). *Big Data Affirmative Action* (SSRN Scholarly Paper 4024623).
<https://papers.ssrn.com/abstract=4024623>
- Samuel, S. (2024, febrero 28). *Black Nazis? A woman pope? That's just the start of Google's AI problem*. Vox.
<https://www.vox.com/future-perfect/2024/2/28/24083814/google-gemini-ai-bias-ethics>
- Silverio, M. (2024). *ChatGPT: Usuarios y estadísticas (2024)*. <https://www.primeweb.com.mx/chatgpt-usuarios-estadisticas>
- Sorbán, K. (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, P. (2017, junio 27). *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>
- Wachter, S., Mittelstadt, B., & Russell, C. (2024). *Do large language models have a legal duty to tell the truth?* (SSRN Scholarly Paper 4771884). <https://doi.org/10.2139/ssrn.4771884>
- Yang, Z. (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, J. C. (2022). *Silicon values: The future of free speech under surveillance capitalism*. Verso Books.
<https://books.google.com/books?hl=es&lr=&id=UOhOEAAAQBAJ&oi=fnd&pg=PR7&dq=Jillian-York->

Silicon-Values_-The-Future-of-Free-Speech-Under-Surveillance-Capitalism-_2021&ots=Q-3sQMADr1&sig=WABVExBL15tjgaaMi7X0M_e5XTY