IDP REVISTA D'INTERNET, DRET I POLÍTICA
REVISTA DE INTERNET, DERECHO Y POLÍTICA

https://idp.uoc.edu

Monográfico: «Digitalización y algoritmización de la justicia» (coord.: F. Miró)

¿Sueña ChatGPT-4 con tweets ofensivos? Una aproximación a las contribuciones potenciales de los modelos generativos en la detección de discurso ilícitos

Mario Santisteban Galarza
Universidad del País Vasco

Jesús C. Aguerri
Centro Crímina

Fecha de recepción: junio 2023
Fecha de aceptación: julio 2023
Fecha de publicación: noviembre 2023

Resumen

Los últimos avances en inteligencia artificial generativa parecen permitir dotar a los modelos de inteligencia artificial de capacidades tan relevantes para el ámbito jurídico como la de argumentar sus propias decisiones. Este estudio se aproxima a las capacidades del modelo ChatGPT-4 en el contexto de la detección de discursos de odio según su conceptualización en el artículo 510.1. a) del Código Penal español. Para ello, se compararán los razonamientos y las decisiones de ChatGPT-4 a partir de relatos de hechos probados con las decisiones de los órganos que juzgaron los respectivos casos, estudiando los límites y el potencial de estos sistemas en el ámbito jurídico.

Palabras clave

inteligencia artificial; discurso del odio; ChatGPT; inteligencia artificial generativa; derechos fundamentales

Does ChatGPT-4 dream of offensive tweets? An approximation to the potential contributions of generative models in detecting illicit speeches

Abstract

The latest advances in generative artificial intelligence appear to allow artificial intelligence models to be equipped with capabilities as relevant to the legal field as to arguing their own decisions. This study approximates the capabilities of the ChatGPT-4 model in the context of the detection of hate speeches according to their conceptualization in article 510.1. a) of the Spanish Criminal Code. For this purpose, ChatGPT-4's reasoning and decisions will be compared from proven factual accounts with the decisions of the bodies that judged the respective cases, studying the limits and potential of these systems in the legal field.

Keywords

artificial intelligence; hate speech; ChatGPT; generative artificial intelligence; fundamental rights

Introducción

La inteligencia artificial (IA) está cada vez más presente en nuestras sociedades, constituyendo tanto un riesgo como una oportunidad para la realización material de la dimensión objetiva de los derechos fundamentales (Presno Linera, 2022). De ahí que las discusiones actuales sobre esta materia se estructuren alrededor de posiciones encontradas, que exacerban las consecuencias negativas o positivas de la tecnología (Miró Llinares, 2022). Por su lado, la posición del regulador europeo, todavía en construcción, acepta gran parte de estas herramientas, si bien estableciendo un sistema de control en función del riesgo especifico que presenta cada una de ellas.[1] En este escenario cambiante, han recibido especial relevancia mediática nuevos modelos generativos de inteligencia artificial, tales como ChatGPT Google Bard o MidJourney. Estos dan cuenta de importantes avances en el campo del procesamiento del lenguaje natural y han mostrado rendimientos muy notables en tareas complejas, como la generación de texto, la respuesta a preguntas o la producción de imágenes.

Tal es la promesa de capacidad para responder a preguntas complejas, de forma aparentemente razonada, que cabe preguntarse si estos nuevos modelos de IA son capaces de emular la argumentación jurídica, con lo que podrían ser utilizados para realizar ciertas tareas que se presuponen a esta última. En esta línea, el presente estudio explora la capacidad del modelo generativo ChatGPT-4 en el marco de la detección de discursos de odio reprimidos por el artículo 510.1. a) del Código Penal español (en lo que sigue, CP). Empujando a ChatGPT a tomar una decisión sobre el fondo, considerando los hechos probados de una muestra de sentencias que utilizan dicho fundamento jurídico, se compara la decisión final adoptada por la IA (absolución/condena) con la decisión adoptada por el Tribunal en el caso real, así como los argumentos esbozados por ChatGPT.

1. La detección de discursos ilícitos en el ciberespacio y la IA generativa

Los sistemas de inteligencia artificial se encuentran cada vez más presentes en el sistema de justicia, obligando a replantear procedimientos jurídicos como las bases de nuestros razonamientos epistémicos (Castro-Toledo *et al*., 2023). En múltiples países ya se han implementado herramientas de valoración del riesgo que asisten al juez y a otros profesionales a la hora de tomar ciertas decisiones, tales como la valoración de la peligrosidad, de riesgo de incumplimiento de ciertas medidas, etc. (Andrés-Pueyo *et al*., 2017; Martínez-Garay, 2018; Quijano-Sánchez *et al*., 2018). El recurso a estos sistemas, de momento de carácter bastante básico, es todavía limitado. Sin embargo, la inteligencia artificial es ampliamente utilizada por plataformas digitales para eliminar contenido contrario a sus condiciones, que abarcan también el discurso de odio (Duarte y Llansó, 2017; Gorwa *et al*., 2020).

Los sistemas de moderación algorítmica han sido criticados por no atender a las complejidades del lenguaje y al contexto cultural en el que se inserta (Dias Oliva *et al*., 2021; Udupa *et al*., 2022). Diversos autores también temen que puedan calificarse de facto como una censura previa (Llansó, 2020), y que puedan limitar la visibilidad de los contenidos antes de que los usuarios lleguen si quiera a interactuar con ellos. En la misma línea, se han percibido como una amenaza para la libertad de expresión en el marco internacional (Kaye, 2018), y el propio Tribunal de Justicia de la Unión Europea ha resaltado que constituyen una limitación de tal derecho, únicamente admisible si la legislación que empuja a su introducción establece las garantías adecuadas (STJUE, de 26 de abril de 2022, asunto C-401/19). Sin embargo, pese a sus limitaciones, lo cierto es que las plataformas apuestan por su implementación al considerar que el volumen de mensajes que circula en sus servicios es inabarcable para una revisión manual (Bloch-Wehba, 2020; Gillespie, 2020).

Recientes normas europeas no vedan la moderación algorítmica sin tampoco imponerla, con alguna excepción en el ámbito de la protección de los derechos de la propiedad intelectual.[2] Ante el riesgo que estos acarrean, el reciente Reglamento 2022/2065, de 19 de octubre de 2022, relativo a un mercado único de servicios digitales, se ha preocupado de ofrecer una tutela amplia frente a estos sistemas. En lo que interesa a efectos de la investigación, establece

1. Nos referimos a la propuesta de reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión (COM/2021/206 final).
2. Nos referimos al discutido artículo 17 de la Directiva (UE) 2019/790 del Parlamento europeo y del Consejo de 17 de abril de 2019 sobre los derechos de autor y derechos afines en el mercado único digital y por la que se modifican las Directivas 96/9/CE y 2001/29/CE.

obligaciones de transparencia frente a estos (art. 15.1. e), y obliga a ofrecer una explicación sobre los motivos en los que se fundamentan las actividades de moderación algorítmica; en el caso de contenidos ilícitos, «una referencia al fundamento jurídico utilizado y explicaciones de por qué la información se considera contenido ilícito conforme a tal fundamento», y en el caso de contenidos contrarios a los términos y condiciones de los intermediarios, «una referencia al fundamento contractual utilizado y explicaciones de por qué la información se considera incompatible con tal fundamento» (art. 17.3). Estas previsiones se acompañan con la obligación de tener en consideración los intereses legítimos de los usuarios y sus derechos fundamentales (art. 14.4) en los procesos de moderación de contenidos. Todo ello obliga a que las plataformas que utilizan sistemas de moderación de contenidos ofrezcan una explicación a los usuarios afectados por las medidas tomados por sus sistemas automatizados, explicaciones que por el momento suelen ser escasamente desarrolladas.

Estas obligaciones legales podrían verse colmadas por los últimos avances en materia de inteligencia artificial, que están llegando de la mano de los denominados modelos generativos. Su principal característica es que ofrecen resultados óptimos a la hora de crear contenido, e «imitar» el lenguaje y el razonamiento humanos. Estos modelos utilizan técnicas de aprendizaje no supervisado para encontrar estructuras y patrones dentro de los datos con los que son entrenados (generalmente texto e imágenes), estableciendo relaciones entre ellos de carácter probabilístico, y de este modo son capaces de generar nuevos datos, es decir, nuevas imágenes, textos, sonidos, etc. Dentro de los modelos generativos, los modelos de lenguaje a gran escala (*Large Lenguaje models*) están trayendo notables avances en el campo del procesamiento del lenguaje natural (o *Natural Lenguaje Processing*), ya que presentan buenos resultados «comprendiendo» el lenguaje humano y produce respuestas acorde a las instrucciones dadas (Ouyang *et al*., 2021), lo que convierte a este tipo de sistemas en herramientas muy útiles para crear chatbots, es decir, sistemas con los que el usuario puede interaccionar en su propio lenguaje (Crespo Miguel y Domínguez Cabrera, 2020).

Este es el caso de las diferentes iteraciones de ChatGPT. Como lo definen sus creadores, la empresa Open AI (2023b), ChatGPT es un modelo entrenado para recibir una instrucción a través de un *prompt* –un texto introducido por el usuario, aunque ChatGPT-4 ya puede recibir instrucciones en forma de imágenes– y generar una respuesta detallada. No es más que un modelo capaz de recibir imágenes o textos como *inputs* y generar textos como *outputs* (Ouyang *et al*., 2021), lo cual hace de forma fundamentalmente probabilística, es decir, determinando cuál es la palabra que con mayor probabilidad seguirá a una palabra previa.

Dada la gran capacidad de los modelos de lenguaje a gran escala para redactar textos y entender instrucciones complejas, sus potencialidades son enormes, hasta el punto de poder convertirse en herramientas de gran utilidad en múltiples sectores. El último modelo presentado, ChatGPT-4, todavía se encuentra en fase beta, pero ya es capaz de superar con muy buenos resultados exámenes de nivel universitario en el sistema norteamericano (OpenAI, 2023a). En este sentido, es difícil sostener que el ámbito jurídico pueda mantenerse indiferente a este tipo de herramientas. ChatGPT-3 ya ha sido usado para explorar las posibilidades del uso de la IA generativa en el campo legal y sus propias limitaciones, con resultados sofisticados pero limitados a tareas en las que hay profesionales que pueden ejercer un control sobre la IA (Perlman, 2022). En este sentido, se ha apuntado que ChatGPT-3 puede ayudar a aliviar las cargas de profesores de Derecho, más allá de labores mecánicas, al redactar, por ejemplo, una carta de recomendación. No obstante, pese a su potencial para redactar exámenes y otros materiales, es frecuente que entre sus respuestas se encuentren errores u omisiones de datos relevantes (Pettinato, 2023), pues como recuerda OpenAI –y a menudo el propio modelo– ChatGPT-4 es un mero modelo de lenguaje, no ha recibido entrenamiento especifico en cuestiones jurídicas y, por tanto, no puede sustituir a un experto. De ahí que su uso en el ámbito judicial haya despertado controversias (Parikh et al., 2023). A pesar de ello, la discusión entre las limitaciones y potencialidades de esta clase de sistemas está muy lejos de cerrarse, tanto por la juventud y complejidad del ámbito como por el avance casi constante de estas herramientas.

2. El discurso de odio en la jurisprudencia nacional

Una serie de contenidos que son sistemáticamente reprimidos tanto por los ordenamientos jurídicos nacionales como por las plataformas digitales son aquellos que se identifica con el llamado discurso de odio. Este permite múltiples

definiciones, pero conforme a una reciente recomendación del Consejo de Ministros del Consejo de Europa este puede entenderse, de una manera amplia, como «todo tipo de expresión que incite, promueva, difunda o justifique la violencia, el odio o la discriminación contra una persona o grupo de personas, o que los denigre, por razón de sus características o estatus personales, reales o atribuidos, como la «raza», el color, la lengua, la religión, la nacionalidad, el origen nacional o étnico, la edad, la discapacidad, el sexo, la identidad de género y la orientación sexual».[3]

En el panorama europeo, el Tribunal Europeo de Derechos Humanos ha manifestado que los discursos políticos que incitan al odio basado en prejuicios religiosos, étnicos o culturales representan un peligro para la paz social y la estabilidad política, y en consecuencia su represión penal puede suponer una interferencia admisible en una libertad básica como es la libertad de expresión (STEDH, de 16 de julio de 2009, caso Feret vs. Bélgica, para. 70), incluso encontrándose en ocasiones *extra muros* de esta con apoyo en al artículo 17 del CEDH (Teruel Lozano, 2017).

Si descendemos al plano de nuestro ordenamiento, cuando nos referimos al discurso de odio lo hacemos en el marco del Derecho penal español, y concretamente a los delitos contemplados en el artículo 510 del Código Penal (CP). Si nos centramos en el artículo 510.1, en su letra a), este castiga a quienes públicamente fomenten, promuevan o inciten directa o indirectamente al odio, hostilidad, discriminación o violencia contra un grupo, una parte de este o contra una persona determinada por razón de su pertenencia a aquel, por motivos racistas, antisemitas, antigitanos u otros referentes a la ideología, entre otros.[4]

Según el Tribunal Supremo, «El elemento nuclear del hecho delictivo consiste en la expresión de epítetos, calificativos, o expresiones, que contienen un mensaje de odio que se transmite de forma genérica» (STS 396/2018, de 9 de febrero de 2018). Desgranemos brevemente y sin ánimo de exhaustividad los elementos del tipo.

En cuanto al sujeto pasivo, existe consenso doctrinal y jurisprudencial en que los delitos del artículo 510 protegen a ciertos «grupos diana», y así el odio solo tiene virtualidad penal en nuestro sistema cuando la hostilidad se dirige a una serie de colectivos protegidos. De ahí que pueda apuntarse a que el odio equivalga a una suerte de «aversión discriminatoria» (Fuentes Osorio, 2021), idea que respaldan instrumentos supranacionales.[5] Consiguientemente, se ha defendido que el bien jurídico protegido por el artículo 510 sea la protección de la igualdad y la prohibición de no discriminación, conectadas con la dignidad de los colectivos afectados.[6] El TS ha señalado que la «vulnerabilidad» del colectivo no es un elemento integrante del tipo (STS 1644/2022, de 5 de mayo), si bien reciente jurisprudencia del propio Tribunal parece contradecir este razonamiento.[7] En cualquier caso, por mor del principio de taxatividad de la ley penal y de la prohibición de analogía contra reo, parece claro que los colectivos enumerados en el artículo 510 deben entenderse como numerus clausus, «no siendo posible su aplicación a otros distintos».[8] Esta idea se encuentra presente en la citada STS 1644/2022, de 5 de mayo, y de forma nítida en la STS 1404/2023, de 11 de abril, en la que se afirma que «el precepto en todo caso extiende su ámbito de protección sobre los grupos que se detallan en el mismo, o las personas que pertenezcan a ellos».

3. Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech (Adopted by the Committee of Ministers on 20 May 2022 at the 132nd Session of the Committee of Ministers).
4. El artículo 510 contempla distintos tipos penales cuyo estudio no abarcará este trabajo. Para ello, véase Circular 7/2019, de 14 de mayo, de la Fiscalía General del Estado, sobre pautas para interpretar los delitos de odio tipificados en el artículo 510 del Código Penal y Alastuey Dobón, 2016.
5. En esta línea la Decisión Marco 2008/913/JAI del Consejo, de 28 de noviembre de 2008 relativa a la lucha contra determinadas formas y manifestaciones de racismo y xenofobia mediante el derecho penal define el concepto de odio como el odio basado en la raza, el color, la religión, la ascendencia o el origen nacional o étnico.
6. Circular 7/2019, de 14 de mayo, de la Fiscalía General del Estado, sobre pautas para interpretar los delitos de odio tipificados en el artículo 510 del Código Penal. También el TS ha señalado que «La esencia de lo que se trata de proteger con este delito ubicado en el artículo 510 CP está en la prohibición de la discriminación, como derecho autónomo derivado del derecho a la igualdad» (STS 1644/2022, de 5 de mayo).
7. Hay que recalcar que, previamente, el TS había señalado, si bien *obiter dicta*, que «El bien jurídico protegido por el tipo penal del artículo 510 es la dignidad de las personas, y colectivos de personas, a los que por su especial vulnerabilidad el Código otorga una protección específica en el mencionado artículo», aseveración que contradice el pronunciamiento más reciente (STS 47/2019, de 4 de febrero).
8. Circular 7/2019, de 14 de mayo, de la Fiscalía General del Estado, sobre pautas para interpretar los delitos de odio tipificados en el artículo 510 del Código Penal.

Respecto al elemento objetivo, el Tribunal Supremo configura el artículo 510 del CP como un delito que no requiere un resultado lesivo (STS 2085/2022, de 19 de mayo), y ha reiterado que «El tipo penal requiere para su aplicación la constatación de la realización de unas ofensas incluidas en el discurso del odio pues esa inclusión ya supone la realización de una conducta que provoca, directa o indirectamente, sentimientos de odio, violencia, o de discriminación» (STS 72/2018, de 9 de febrero). Son, en palabras del Tribunal, «expresiones que, por su gravedad, por herir los sentimientos comunes a la ciudadanía, se integran en la tipicidad» (STS 72/2018, de 9 de febrero), debiendo recordarse que el mencionado precepto no castiga únicamente la incitación a cometer hechos delictivos contra los colectivos dianas, sino la incitación al propio odio contra esos colectivos (Alastuey Dobón, 2016). Así, el TS ha defendido que el artículo 510.1 es un delito de peligro abstracto (STS 72/2018, de 9 de febrero; STS 4283/2020, de 11 de diciembre; STS 2085/2022, de 19 de mayo), no debiéndose de «entrar en disquisiciones sobre si la provocación ha de ser directa o indirecta» (STS 4283/2020, de 11 de diciembre). Esto puede chocar con ciertos pronunciamientos del Tribunal Constitucional, que exigió al menos la incitación indirecta para realizar una distinción entre la mera difusión de ideas y la realización de conductas expresivas lesivas de derechos individuales o colectivos (STC 235/2007). En este sentido, se ha criticado que incluso la aceptación de una incitación indirecta supone ya la imposibilidad de diferenciar entre la difusión de ideas y una vejación penalmente reprobable (Rodríguez Montañés, 2012). En todo caso, y sin ser este el fin de este trabajo, cabe apuntar que sin exigirse la posibilidad real de un daño material y adentrándonos en el terreno de la legitimación del castigo penal de comportamientos meramente ofensivos (Feinberg, 1985), resulta como mínimo cuestionable que la lucha contra estos discursos se realice a través de la imposición de penas privativas de libertad (Miró Llinares, 2015).

Finalmente, respecto al elemento subjetivo, «El dolo de estos delitos se rellena con la constatación de la voluntariedad del acto y la constatación de no tratarse de una situación incontrolada o una reacción momentánea, incluso emocional, ante una circunstancia que el sujeto no ha sido capaz de controlar» (STS 396/2018, de 9 de febrero). De ahí que «para afirmar el dolo es suficiente el conocimiento por el autor de los elementos que definen el tipo objetivo, esto es, la plena conciencia y voluntad de que se está difundiendo un mensaje de odio en el que se menosprecia todo aquello que resulta no aceptado por la particular visión del mundo y de la vida que suscribe el emisor y en los que se invita a luchar contra el "enemigo" recurriendo, si resulta preciso, a la violencia» (STS 2085/2022, de 19 de mayo).

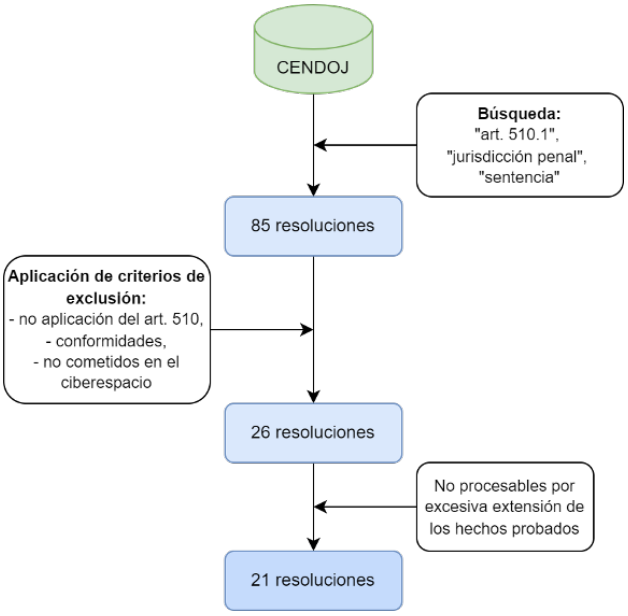
3. ChatGPT-4 y la detección y enjuiciamiento de discurso del odio: una aproximación empírica

El presente estudio busca aproximarse a las capacidades potenciales de los nuevos modelos generativos en la detección de discursos problemáticos. Admitiendo que el proceso de valoración de lo que constituye un discurso problemático es un proceso eminentemente subjetivo, y en consecuencia difícilmente parametrizable en términos de error o acierto (Gillespie, 2020), se optó por acudir a ejemplos jurisprudenciales con los que medir el rendimiento del sistema en términos de concordancia con el fallo y las argumentaciones del juzgador. Para ello se creó una base de datos de sentencias provenientes de órganos jurisdiccionales del orden jurisdiccional penal, que versan sobre el mencionado artículo 510.1 del Código Penal. Dichas sentencias se obtuvieron a través del buscador del Centro de Documentación Judicial (CENDOJ). Concretamente, se realizó una búsqueda con los siguientes parámetros: término de búsqueda «art. 510.1. CP», acotando la jurisdicción a «jurisdicción penal» y el tipo de resolución a «sentencia». Los resultados del buscador arrojaron 85 resoluciones judiciales, que fueron descargadas, de las cuales una de ellas se extravió por un error en el acceso al formato, quedando 84 sentencias para su posterior análisis.

Una vez realizado esta búsqueda inicial, se procedió a la lectura individual de las sentencias, realizando un segundo cribado. En primer lugar, se eliminaron aquellas decisiones que pese a contener la referencia al artículo 510.1 del CP, no adoptaban una decisión sobre el fondo apoyándose en el precepto, versando sobre la aplicación de otros tipos penales. Posteriormente, se prescindió de aquellas sentencias que se limitaban a ratificar la conformidad de acusación y defensa sobre la calificación jurídica de los hechos, debido a que en estos casos el juez o tribunal no realizaba una valoración jurídica de los hechos probados. Por último, se prescindió también de aquellas sentencias que no valoraban la licitud de comentarios o contenidos difuminados mediante servicios de la sociedad de la información.

Una vez finalizado el proceso de cribado, la base de datos quedó constituida por 26 sentencias de distintos órganos jurisdiccionales penales, entre los que se encuentran el TS y otros tribunales, como los tribunales superiores de Justicia y las audiencias provinciales. De estas 26 se extrajeron los hechos probados considerados para realizar la correspondiente valoración sobre la tipicidad de las conductas. Estos fueron presentados a Chat GPT 4, encabezados con la siguiente tarea: «Actuando como un juez de lo penal que ejerce en un juzgado español, determina si los siguientes hechos probados pertenecientes a diferentes casos son constitutivos de un delito de odio encuadrable en el artículo 510.1 del Código Penal español. En consecuencia, determina si se debe absolver o condenar al acusado y razona en términos jurídicos la decisión, escribiendo un ensayo que explique el porqué de la condena o la absolución que figurará en la sentencia».[9] De las 26 sentencias que ChatGPT-4 analizó, 5 tuvieron que ser descartadas al presentar un texto demasiado largo para la herramienta.[10]

Figura 1. Proceso de selección de sentencias



Fuente: elaboración propia

De las restantes 21 sentencias, en su juicio ChatGPT-4 coincidió en 17 de ellas con la decisión original del juzgador. En concreto, en aquellas decisiones que resultaron en una condena (7 sentencias), ChatGPT-4 también acabó por entender los hechos probados como constitutivos del artículo 510.1 del CP. Por el contrario, de las 14 resoluciones restantes, siendo todas ellas absolutorias conforme al artículo 510.1 del CP, ChatGPT-4 entendió que los hechos probados de 4 de ellas eran constitutivos de un delito encuadrable en el artículo 510.1 del CP, mientras que los 10 restantes valoraron que los hechos probados no merecían reproche penal. Estos son ciertamente positivos, ya que la IA acertó en un 81 % de las ocasiones, una exactitud bastante buena pese a que la tendencia de la IA a condenar en casos en los que los jueces absolvieron llegue a lastrar la precisión del modelo.

Tabla 1. Decisiones adoptadas

		Órgano Judicial	
		Absolución	Condena
ChatGPT-4	Absolución	10	0
	Condena	4	7

Fuente: elaboración propia

Al adentrarnos en las respuestas del sistema en la resolución de los casos, observamos algunas carencias. Dado que el modelo solo cuenta con información previa a septiembre de 2021, no incluye en ningún momento la discriminación antigitana o la aporofobia, motivos introducidos por la Ley Orgánica 6/2022, de 12 de julio. Más aún, en uno de sus razonamientos ChatGPT-4 menciona a grupos que no pueden constituir sujeto pasivo del mencionado artículo 510.1. Otro error reseñable es que en uno de sus pronunciamientos utilizó parte de la redacción del antiguo artículo 510.2 del CP, y mezcló el elemento del desprecio a la verdad con el de promoción o incitación directa o indirecta al odio, ampliando la redacción de tipo.[11]

9. Somos conscientes de que el artículo 510.1 del CP no abarca una única conducta típica, sino que en sus distintos numerales se prevén tipos específicos. No obstante, en la jurisprudencia que conforma la base de datos, y utilizada para medir el rendimiento de ChatGPT-4, no se realiza habitualmente tal distinción, por lo que decidimos, a riesgo de imprecisión, ofrecer a la herramienta un margen más amplio a la hora de seleccionar las conductas que abarca el mencionado artículo 510.1.
10. Al encontrarse el sistema en fase de pruebas, la empresa responsable Open AI solo ofrece de forma abierta acceso a ChatGPT-4 mediante su página web, en la que se ha impuesto un límite de palabras a las interacciones con el modelo, sin embargo, si se accede a través de API a él –una opción actualmente en beta cerrada–, ChatGPT-4 es capaz de procesar 32.768 tókenes, lo que equivale a unas 30.000 palabras.
11. En este caso, ChatGPT-4 añade también el verbo «provocar» al tipo, lo cual, además de no constar en la redacción del vigente artículo 510.1. a), puede incitar a confusión con el artículo 18 del CP.

Por otra parte, si bien ChatGPT-4 responde a la solicitud de valoración realizada en los casos señalados sus respuestas son por lo general escuetas,[12] y no ofrece información jurídica relevante para apoyar sus decisiones más allá del citado precepto legal. En consecuencia, y coincidiendo con las opiniones de otros expertos legales que se han acercado al sistema, así como con las recomendaciones de la propia Open AI, debe destacarse la poca confiabilidad del sistema para ofrecer asistencia jurídica al público general sin una supervisión de un profesional.

A pesar de las limitaciones apuntadas deben destacarse algunos aspectos notablemente positivos de las respuestas de la IA. En general, ChatGPT-4 se refiere a los hechos probados, identificándolos adecuadamente y utilizándolos en su «argumentación», lo que da cuenta de su gran capacidad para comprender, procesar y elaborar textos. Por ejemplo, juzgando los hechos de la STS 2085/2022, de 19 de mayo, el modelo sintetiza así los hechos probados:

> «Los hechos probados muestran que el acusado, Antonio, también conocido como Argimiro, ha utilizado diversos perfiles en una plataforma de comunicación para promover y difundir sentimientos de odio, discriminación y violencia contra diferentes colectivos, incluyendo otras ramas del Islam, el pueblo español en general, los andaluces en particular, y personas con distintas orientaciones sexuales. Además, ha publicado mensajes degradantes sobre las mujeres».

Asimismo, y pese a ser reducidas en cuanto a su extensión, los veredictos de la IA se apoyan adecuadamente en los hechos probados, poniendo estos en relación con el tipo y apoyándose en valoraciones relativamente complejas. A efectos ilustrativos, cabe mencionar el veredicto de ChatGPT-4 sobre los hechos de la SAP de Barcelona 7610/2019, de 21 de mayo:

> «En cuanto a la provocación, promoción o incitación al odio, hostilidad, discriminación o violencia, el comentario de Ruperto expresa un deseo de daño hacia el Sr. Juan Pablo basado en su origen nacional. Si bien el comentario es ofensivo y despectivo, no se desprende de él una incitación directa a otros a cometer actos de odio, hostilidad, discriminación o violencia. Sin embargo, la incitación indirecta también es sancionada por el art. 510.1 del Código Penal, y en este caso, podría entenderse que el comentario de Ruperto fomenta la discriminación y el rechazo hacia personas de origen marroquí, lo cual podría constituir una incitación indirecta al odio».

Destacan también varias argumentaciones realizadas al hilo de pronunciamientos sobre hechos probados de sentencias absolutorias. ChatGPT-4 coincide en dos ocasiones con el juzgador en la absolución basándose en que el sujeto pasivo afectado por las expresiones enjuiciadas no es un colectivo previsto en el artículo 510.1. a). En el primer caso, ChatGPT-4 apuntó que «Aunque algunos de los tuits pueden ser considerados ofensivos o inapropiados, no se ha demostrado que inciten al odio, la hostilidad, la discriminación o la violencia contra un grupo o persona determinada por razón de su pertenencia a ese grupo», dudando igualmente la Audiencia Provincial que el colectivo de «no independentistas» estuviese protegido por el artículo 510.1 del CP.[13] Más claro es el supuesto que se enjuició en la SAP de Segovia 38/2020, de 13 de marzo. En él, la Audiencia Provincial entendió que el colectivo taurino no se encontraba protegido por el artículo 510.1 del CP defendiendo la atipicidad de la conducta. ChatGPT-4 tomó la misma decisión, que argumentó en los siguientes términos:

> «En este caso, los mensajes del acusado están dirigidos a toreros y no a un grupo específico protegido por el artículo 510.1 del Código Penal. Aunque los comentarios son ofensivos, no parece que promuevan el odio, la hostilidad, la discriminación o la violencia basados en alguno de los motivos protegidos por dicho artículo».

12. Dado que el sistema se encuentra en beta, Open AI limita la extensión de las respuestas, configurando internamente el modelo para que tienda a ser sintético. Sin embargo, cuando el acceso a través de API esté disponible, el usuario podrá configurar este parámetro a voluntad, como ya ocurre con los modelos previos de Open AI.
13. «En el presente caso, surge la primera duda razonable para este Tribunal siendo esta que entendemos el grupo "no independentista" estaría formado por todas aquellas personas que están de acuerdo con el orden constitucional vigente, y que por tanto comulgan con las estructuras del Estado y con la protección que otorga a través del funcionamiento de sus instituciones, sin que de la misma se desprenda que se encuentren en "minoría" ni siquiera en Cataluña (no está claro cuál es la minoría, si es cuantitativa o cualitativa, en ese efecto de hacerse oír o hacer llegar el mensaje circunscrito a un ámbito territorial) y que necesiten de una especial protección, más allá que la que otorgan las instituciones» (SAP B 14641/2018).

Por último, también cabe destacar que en algunos casos el propio ChatGPT-4 inserta la ponderación de la libertad de expresión en la discusión. Así, considerando los hechos probados de la SAP de Tenerife 1838/2021, de 1 de junio de 2016, la IA señala que: «gran parte de las publicaciones parecen estar enfocadas en la crítica al gobierno de Israel y a sus políticas en relación con el conflicto con Palestina, más que en incitar al odio o la discriminación hacia un grupo específico por razones de religión, ideología u origen nacional», criterio coincidente con el de la Audiencia, si bien acompañado de un razonamiento mucho más rico.

4. Discusión: *eppur si muove*

ChatGPT no está diseñado para la toma de decisiones judiciales. Es una IA muy versátil capaz de generar contenido y replicar la comunicación humana, lo que implica no solo procesar instrucciones y generar textos coherentes, sino también simular cierto razonamiento. En consecuencia, el modelo ofrece resultados positivos en aquello para lo que ha sido creado, pero, obviamente, presenta deficiencias evidentes en campos en los que no ha sido entrenado. Es por ello por lo que se desaconseja su uso para realizar decisiones informadas en el campo legal, en línea con otros estudios.

El presente estudio ha tratado de aproximarse a las capacidades de estos sistemas para la detección y el enjuiciamiento de discurso del odio. Este sufre limitaciones, como el número de resoluciones que han podido ser consideradas para juzgar las capacidades de ChatGPT-4 al utilizar únicamente la base de datos CENDOJ, o la aplicación de criterios de exclusión, como la necesidad de que el delito de odio enjuiciado se cometiese en el ciberespacio. Más aún, el estudio ha adoptado una posición indiferente frente al contenido de las sentencias utilizadas para medir el rendimiento de ChatGPT, presumiendo su corrección con lo establecido en el punto 3, obviando el debate de fondo sobre cuáles son los estándares que el Derecho penal debe seguir en la represión del discurso.

Sin embargo, los resultados que arroja el estudio no dejan de ser prometedores. Las respuestas de la IA de Open AI contienen argumentos que también podemos encontrar en los fundamentos de derecho de las sentencias de las que se extrajeron los hechos probados. La ratio de acuerdo entre decisiones de ChatGPT-4 y los distintos tribunales seleccionados es positiva, pero lo son más ciertas argumentaciones, como la exclusión de la protección de ciertos colectivos por el artículo 510.1, al no encontrarse previstos en la norma. Al fin y al cabo, quizás, en la práctica, la aplicación del derecho no acabe siendo muy distinta de la simulación de un razonamiento basado en unos principios que una vez se estudiaron y en la interpretación de la voluntad desconocida tras los textos en los que se ha sido entrenado. Si atendemos a los resultados preliminares de estas herramientas –las algorítmicas–, repetimos, no entrenadas para operar específicamente en este campo, su uso es una posibilidad que no debe descartarse.

Ahora bien, si esto aún parece lejano, su adecuación para otras labores, como la revisión automatizada de contenidos es más factible. Cierta automatización parece ineludible en el escenario actual, por lo que sistemas que puedan realizar una «valoración» de los hechos, llegar a una decisión y fundamentarla de cara al usuario se hacen necesarios para cumplir con las exigencias de los recientes textos legales europeos, entre ellos el Reglamento de Servicios Digitales. La valoración exigida en estos procesos no debe revestir la calidad que se espera del razonamiento jurídico en el marco de un proceso judicial, que debe ser elevada para respetar los derechos en el proceso y tildarse de garantista. Esto es fácticamente imposible por las condiciones en las que se ejerce la moderación, entre ellas, la celeridad en la respuesta que se exige a estos procesos (Duoek, 2022). A la luz de la menor relevancia jurídica que puede tener la decisión de las plataformas digitales en comparación con la sanción penal, es justamente en estos procesos de moderación de contenidos donde puede plantearse el uso de sistemas generativos.

Los resultados de la presente investigación no pueden ser extrapolados a este concreto sistema de decisión, aunque ambos, moderación de contenidos y la aplicación de la ley penal, consistan en procesos de valoración de conductas y su encaje en ciertas categorías de comportamientos desviados. En el futuro deberán realizarse otros estudios que comparen las capacidades de los sistemas generativos con los existentes sistemas de detección de contenidos ilícitos que se encuentran al alcance de los investigadores, y determinar si efectivamente la IA generativa ofrece buenos resultados en este ámbito.

Reconocimientos

Este trabajo se realiza en el marco del proyecto *Ius_ machinA* (Sobre las bases normativas y el impacto real de la utilización de algoritmos predictivos en los ámbitos judicial y penitenciario) (TED2021- 129356B-I00), financiado por MCIN/AEI/10.13039/501100011033 y la Unión Europea *NextGenerationEU/PRTR*, y asimismo es posible gracias la financiación derivada de la convocatoria PIF 2020 (UPV/EHU) y la FJC2020-042961-I, financiada por el Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación/ 10.13039/501100011033 y la Unión Europea NextGenerationEU/PRTR. También es posible gracias a la ayuda regulada por la convocatoria PIF UPV/EHU 2020.

Referencias bibliográficas

ALASTUEY DOBÓN, C. (2016). «Discurso del odio y negacionismo en la reforma del Código Penal de 2015». *Revista Electrónica de Ciencia Penal y Criminología*, vol. 18-14, págs.1-38 [en línea]. Disponible: http://criminet.ugr.es/recpc/18/recpc18-14.pdf

ANDRÉS PUEYO, A.; ARBACH-LUCIONI, K.; REDONDO, S. (2017). «The RisCanvi: A New Tool for Assessing Risk for Violence in Prison and Recidivism». En: Jay P. Singh, Daryl G. Kroner, J. Stephen Wormith, Sarah L. Desmarais, Zachary Hamilton (eds). *Handbook of Recidivism Risk/Needs Assessment Tools*, págs. 255-268. John Wiley & Sons Ltd. DOI: https://doi.org/10.1002/9781119184256.ch13

BLOCH-WEHBA, H. (2020). «Automation in Moderation». *Cornell International Law Journal*, vol. 53, págs. 42-96 [en línea]. Disponible en: https://community.lawschool.cornell.edu/wp-content/uploads/2021/03/Bloch-Wehba-final.pdf

CRESPO MIGUEL, M.; DOMÍNGUEZ CABRERA, B. (2020). «Perspectivas de las tecnologías de Chatbot y su aplicación a las entrevistas de evaluación del lenguaje». *Pragmalingüística*, vol. 2, págs. 100-113. DOI: https://doi.org/10.25267/Pragmalinguistica.2020.iextra2.06

CASTRO-TOLEDO, F. J.; MIRÓ-LLINARES, F.; AGUERRI, J. C. (2023). «Data-Driven Criminal Justice in the age of algorithms: epistemic challenges and practical implications». *Crim Law Forum*, vol. 34, pags. 295-316. DOI: https://doi.org/10.1007/s10609-023-09454-y

DIAS OLIVA, T.; ANTONIALLI, D. M.; GOMES, A. (2021). «Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online». *Sexuality & Culture*, vol. 25, n.º 2, págs. 700-732. DOI: https://doi.org/10.1007/s12119-020-09790-w

DUARTE, N.; LLANSÓ, E. (2017). *Mixed Messages? The Limits of Automated Social Media Content Analysis. CTD* [en línea]. Disponible en: https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/. [Fecha de consulta: 31 de mayo de 2023].

DUOEK, E. (2022). «Content Moderation as Systems Thinking». *Harvard Law Review*, vol. 136, 528-606. DOI: http://dx.doi.org/10.2139/ssrn.4005326

FEINBERG, J. (1985). *Offense to Others (The Moral Limits of Criminal Law)*, vol. 2. Oxford: Oxford University press.

FUENTES OSORIO, J. L. (2021). «El odio como delito». *Revista Electrónica de Ciencia Penal y Criminología*, vol. 19-27, págs. 1-52 [en línea]. Disponible en: http://criminet.ugr.es/recpc/19/recpc19-27.pdf

GILLESPIE, T. (2020). «Content moderation, AI, and the question of scale». *Big Data & Society*, vol. 7, n.º 2. DOI: https://doi.org/10.1177/2053951720943234

GORWA, R.; BINNS, R.; KATZENBACH, C. (2020). «Algorithmic content moderation: Technical and political challenges in the automation of platform governance». *Big Data & Society*, vol. 7, n.º 1. DOI: https://doi.org/10.1177/2053951719897945

HELBERGER, N.; DIAKOPOULOS, N. (2023). «ChatGPT and the AI Act». *Internet Policy Review*, vol. 12, n.º 1. DOI: https://doi.org/10.14763/2023.1.1682

HUMAN RIGHTS WATCH (2020). «"Video Unavailable". Social Media Platforms Remove Evidence of War Crimes». *Human Rights Watch* [en línea]. Disponible en: https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes. [Fecha de consulta: 31 de mayo de 2023].

LLANSÓ, E. J. (2020). «No amount of "AI" in content moderation will solve filtering's prior-restraint problem». *Big Data & Society*, vol. 7, n.º 1. DOI: https://doi.org/10.1177/2053951720920686

MARTÍNEZ-GARAY, L. (2018). "Peligrosidad, algoritmos y due process: El caso State vs. Loomis». Revista de Derecho Penal y Criminología, n.º 20, págs. 485-502. DOI: https://doi.org/10.5944/rdpc.20.2018.26484

MIRÓ LLINARES, F. (2015). «La criminalización de conductas "ofensivas". A propósito del debate anglosajón sobre los "límites morales" del derecho penal». *Revista electrónica de ciencia penal y criminología*, n.º 17 [en línea]. Disponible en: http://criminet.ugr.es/recpc/17/recpc17-23.pdf

MIRÓ LLINARES, F. (2022). «Inteligencia artificial, delito y control penal: nuevas reflexiones y Algunas predicciones sobre su impacto en el derecho y la justicia penal». *El Cronista del Estado Social y Democrático de Derecho*, n.º 100, págs. 174-183.

MIRÓ-LLINARES, F.; MONEVA, A.; ESTEVE, M. (2018). «Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments». *Crime Science*, vol. 7, n.º 1. DOI: https://doi.org/10.1186/s40163-018-0089-1

OPENAI (2023a). «GPT-4 Technical Report». *arXiv*. DOI: https://doi.org/10.48550/arXiv.2303.08774

OPENAI (2023b). «Introducing ChatGPT». *OpenAI Blog* [en línea]. Disponible en: https://openai.com/blog/chatgpt#OpenAI. [Fecha de consulta: 21 abril 2023].

OUYANG, L.; WU, J.; JIANG, X.; ALMEIDA, D.; WAINWRIGHT, C. L.; MISHKIN, P.; ZHANG, C.; AGARWAL, S.; SLAMA, K.; RAY, A.; SCHULMAN, J.; HILTON, J.; KELTON, F.; MILLER, L.; SIMENS, M.; ASKELL, A.; WELINDER, P.; CHRISTIANO, P.; LEIKE, J.; LOWE, R. (2021). «Training language models to follow instructions with human feedback». *arXiv*. DOI: https://doi.org/10.48550/arXiv.2203.02155

PARIKH P. M.; SHAH D. M., PARIKH K. P. (2023). «Judge Juan Manuel Padilla Garcia, ChatGPT, and a controversial medicolegal milestone». *Indian Journal of Medical Sciences*, vol. 75, n.º 1, págs. 3-8. DOI: https://doi.org/10.25259/IJMS_31_2023

PETTINATO OLTZ, T. (2023, febrero). «ChatGPT, Professor of Law». *SSRN*. DOI: https://doi.org/10.2139/ssrn.4347630

PERLMAN, A. M. (2022). «The Implications of ChatGPT for Legal Services and Society». *SSRN*. DOI: https://doi.org/10.2139/ssrn.4294197

PRESNO LINERA, M. A. (2022). *Derechos Fundamentales e Inteligencia Artificial*. Madrid: Marcial Pons. DOI: https://doi.org/10.2307/jj.4908196

QUIJANO-SÁNCHEZ, L.; LIBERATORE, F.; CAMACHO-COLLADOS, J.; CAMACHO-COLLADOS, M. (2018). «Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police». *Knowledge-Based Systems*, vol. 149, págs. 155-168. DOI: https://doi.org/10.1016/j.knosys.2018.03.010

RODRÍGUEZ MONTAÑÉS, T. (2012). *Libertad de expresión, discurso extremo y delito. Una aproximación a las fronteras del derecho penal*. Valencia: Tirant lo Blanch.

STOKEL-WALTER, C. (2023). «Generative AI Is Coming for the Lawyers. Large law firms are using a tool made by OpenAI to research and write legal documents. What could go wrong?». *TheWired* [en línea]. Disponible en: https://www.wired.co.uk/article/generative-ai-is-coming-for-the-lawyers. [Fecha de consulta: 31 de mayo de 2023].

TERUEL LOZANO, G. (2017). «El discurso del odio como límite a la libertad de expresión en el marco del convenio europeo». *ReDCE*, n.º 27 [en línea]. Disponible en: https://www.ugr.es/~redce/REDCE27/articulos/03_TERUEL.htm

UDUPA, S.; MARONIKOLAKIS, A.; SCHÜTZE, H.; WISIOREK, A. (2022). «Ethical Scaling for Content Moderation: Extreme Speech and the (In)Significance of Artificial Intelligence». *Big Data & Society*, vol. 10, n.º 1. DOI: https://doi.org/10.1177/20539517231172424. [Fecha de consulta: 31 de mayo de 2023]

Cita recomendada

SANTISTEBAN GALARZA, Mario; AGUERRI, Jesús C. (2023). «¿Sueña ChatGPT-4 con tweets ofensivos? Una aproximación a las contribuciones potenciales de los modelos generativos en la detección de discurso ilícitos». En: Miró, F. (coord.). «Digitalización y algoritmización de la justicia». *IDP. Revista de Internet, Derecho y Política*, núm. 39. UOC [Fecha de consulta: dd/mm/aa]
http://dx.doi.org/10.7238/idp.v0i39.416638

Sobre las autorías

Mario Santisteban Galarza
Universidad del País Vasco
mariosantg@gmail.com

Jurista, investigador predoctoral de la Universidad del País Vasco (UPV/EHU) en el Departamento de Derecho de la Empresa y Derecho Civil. Su trabajo se centra en las relaciones entre el derecho y la tecnología, particularmente en la libertad de expresión en internet y los poderes de las plataformas digitales para controlar el discurso.

Jesús C. Aguerri
Centro Crímina para el Estudio y la Prevención de la Delincuencia de la Universidad Miguel Hernández de Elche
j.aguerri@crimina.es
ORCID: https://orcid.org/0000-0002-7730-8527

Investigador posdoctoral (Juan de la Cierva-Formación en el Centro Crímina para el Estudio y la Prevención de la Delincuencia). Es doctor en Sociología por la Universidad de Zaragoza y ha desarrollado su trabajo alrededor de la sociología del ciberespacio y la gestión de la libertad de expresión.

IDP REVISTA D'INTERNET, DRET I POLÍTICA
REVISTA DE INTERNET, DERECHO Y POLÍTICA

Universitat Oberta de Catalunya