



Descripción de los riesgos y desafíos para la integridad académica de aplicaciones generativas de inteligencia artificial*

Risks and Challenges Posed by Artificial Intelligence Generative Applications for Academic Integrity

ROBERTO NAVARRO-DOLMESTCH**

Universidad Católica del Maule (Chile)

Resumen: Este artículo analiza, desde una perspectiva descriptiva y como apertura de una línea de investigación, el impacto que las tecnologías de inteligencia artificial generativas (IAG) pueden representar para la integridad académica, materializado en la actividad docente y en los procesos de evaluación en la enseñanza universitaria del derecho. El artículo toma como premisa la definición de la integridad académica como un conjunto de valores y sostiene que de la IAG surgen una serie de riesgos que amenazan dichos valores, como la excesiva dependencia y confianza en la IAG, la irrealizabilidad del proyecto pedagógico y la pérdida de competitividad de las instituciones educativas, entre otros. Para minimizar o anular tales riesgos y, de esa forma, impedir que ellos se concreten en afectaciones a la integridad académica, se identifican cuatro medidas de mitigación para ser aplicadas en entornos universitarios.

Palabras clave: Integridad académica, inteligencia artificial, generación automática de texto, *chatbot*, ChatGPT

Abstract: From the perspective of a descriptive analysis, and as a starting point to a new research line, this paper examines the potential impact Generative Artificial Intelligence (GAI) technologies may have on academic integrity, manifested in the learning and evaluation processes of law classes at the university level. The article takes as its premise the definition of academic integrity as a set of values and argues that a series of risks arise from the GAI that threaten those values, such as excessive dependence and trust in the GAI, the unreachability of the pedagogical project and the loss of competitiveness of educational institutions, among others. To minimize or nullify such risks, and thus prevent them from affecting academic integrity, four mitigation measures are identified to be applied in university environments.

Keywords: Academic integrity, artificial intelligence, automatic text generation, *chatbot*, ChatGPT

* Este artículo se ha elaborado en el marco del Proyecto Fondecyt de Iniciación 2023 N.° 11230216, financiado por la Agencia Nacional de Investigación y Desarrollo del Gobierno de Chile, titulado «Aplicación del Derecho Penal a la Inteligencia Artificial», y del que su autor es investigador responsable.

** Doctor en Derecho por la Universidade da Coruña (España). Profesor de Derecho Penal de la Universidad Católica del Maule (Chile).

Código ORCID: 0000-0003-0907-5714. Correo electrónico: ronavarro@ucm.cl

CONTENIDO: I. INTRODUCCIÓN.- II. APLICACIONES DE GENERACIÓN AUTOMÁTICA DE LENGUAJE NATURAL.- III. INTEGRIDAD ACADÉMICA.- IV. CONDUCTAS Y SU (DIS)VALORACIÓN.- V. RIESGOS.- V.1. RIESGOS PROCEDIMENTALES.- V.1.1. EXCESIVA DEPENDENCIA Y CONFIANZA EN LA AIG.- V.1.2. IRREALIZABILIDAD DEL PROYECTO PEDAGÓGICO.- V.1.3. INCORPORACIÓN DE SESGOS Y DISCRIMINACIÓN EN EL APRENDIZAJE.- V.2. RIESGOS DE RESULTADO INDIVIDUALES.- V.2.1. RESPONSABILIDAD LEGAL.- V.2.2. APRENDIZAJE DISTORSIONADO Y ADQUISICIÓN DE SESGOS Y DISCRIMINACIONES.- V.2.3. SOBRECALIFICACIÓN INFUNDADA.- V.3. RIESGOS DE RESULTADO COLECTIVOS.- V.3.1. PÉRDIDA DE COMPETITIVIDAD DE LA INSTITUCIÓN.- V.3.2. CONTAMINACIÓN DE LA CULTURA ACADÉMICA.- V.3.3. DEGRADACIÓN DEL VALOR ÉTICO DE LA INSTITUCIÓN.- VI. MEDIDAS DE MITIGACIÓN.- VI.1. ADAPTACIÓN DE LOS MARCOS NORMATIVOS.- VI.2. REVALORIZACIÓN DEL PENSAMIENTO CRÍTICO.- VI.3. CAMBIOS EN LOS PARADIGMAS DE LA DOCENCIA Y LA EVALUACIÓN.- VI.4. CUESTIONAMIENTO DE MODELOS DE *E-LEARNING*.- VII. CONCLUSIONES.

I. INTRODUCCIÓN

La inteligencia artificial (IA) es un conjunto heterogéneo de tecnologías que pretenden emular las capacidades humanas para el desarrollo de tareas que, hasta su surgimiento, eran propias y exclusivas de los seres humanos (Abbott & Sarch, 2019, p. 329; 2020, p. 179; Barfield & Pagallo, 2020, p. 1; Fierens *et al.*, 2021, p. 51)¹. Tales capacidades se refieren, principalmente, a la percepción del mundo, la comprensión e interpretación de este, el raciocinio, la argumentación y la creatividad. Dentro de este contexto se inscribe la inteligencia artificial generativa (IAG), que es un enfoque tecnológico que consiste en «el aprendizaje por máquinas no-supervisado o semisupervisado para crear nuevo contenido, incluido, pero no limitado, a imágenes digitales, vídeo, audio, texto o código» (Hu, 2022). Los modelos generativos «se definen como métodos o técnicas que son capaces de transformar información en código o un subconjunto de códigos desde el que se puede reconstruir la información» (Abukmeil *et al.*, 2021, p. 2). La IAG se ha desarrollado en torno a dos marcos preferentes: *Generative Adversarial Network* (GAN)

1 En un esfuerzo por entregar una definición más comprensiva de la IA, Russell y Norvig (2020) distinguen cuatro enfoques definitorios, distinción que aún tiene vigencia. El primero, el «enfoque del test de Turing», plantea que la IA es aquella tecnología capaz de interactuar verbalmente con un ser humano sin que este pueda percatarse de que está conversando con una máquina. El segundo, el «enfoque cognitivo», propone que IA es aquella tecnología que piensa como un ser humano en el sentido en que ha sido determinado por la ciencia cognitiva. Se trata de que las máquinas inteligentes resuelvan problemas siguiendo los mismos pasos de razonamiento que un ser humano enfrentado al mismo problema. El tercero, el «enfoque de las leyes del pensamiento», plantea que IA es aquella tecnología capaz de efectuar inferencias correctas; es decir, de aplicar las leyes de la lógica en forma autónoma. En este enfoque, la IA se describe como aquella con la capacidad suficiente de tomar la descripción de un problema expresado en notación lógica y encontrar la solución, si ella existe. Y el cuarto, el «enfoque del agente racional», considera que la IA es aquella capaz de tomar los datos del mundo externo, procesarlos y decidir cursos de acción y soluciones racionales a los problemas que se presenten. La racionalidad implica la correcta construcción de inferencias lógicas; adicionalmente, la racionalidad también se integra de algo que está más allá de las meras inferencias lógicas, para lo cual la IA requiere aplicar conocimientos previos del mundo circundante (pp. 5-8).

(Albahar & Almalki, 2019, p. 3245; Caldera, 2019, p. 181; Raj *et al.*, 2022, p. 181) y *Generative Transformer* (Dale, 2021; Hu, 2022). El primero se puede emplear, por ejemplo, en la generación de imágenes y vídeos digitales, y, con ellos, en usos como las *deepfakes*²; el segundo, por su parte, ha permitido desarrollar los chats conversacionales de generación de contenido como ChatGPT (Generative Pre-trained Transformer) de OpenAI o BERT (Bidirectional Encoder Representations from Transformers) de Google.

Las tecnologías generativas, cuyo desarrollo fue posible gracias al avance en redes neuronales y al aprendizaje automático (*machine learning*), están diseñadas para ejecutar tareas que, generalmente, incluyen resumir o sintetizar textos, traducir automáticamente o generar diálogos, preguntas o paráfrasis; generar texto desde información numérica o estructurada (como tablas) o desestructurada; o generar texto desde multimodalidad, como la conversión de imágenes o vídeos en textos en lenguaje natural (Azaria, 2022; Li *et al.*, 2022) o viceversa. Dentro de esas tecnologías, los Modelos de Lenguaje (ML)³ son aplicaciones informáticas, que pueden estar basadas en IA y que generan contenidos expresados en imágenes digitales (fotografías o vídeos) o en lenguaje natural; esto es, de la misma forma en la que los seres humanos usamos el lenguaje. En este último caso, es difícil o prácticamente imposible para un lector humano determinar si un contenido fue generado por una máquina.

En el caso de ChatGPT, uno de los ML más famosos y difundidos, el usuario puede entablar con la aplicación una interacción que emula

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIAL

RISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

2 El término *deepfake* (traducido como «ultrafalso») por autores como García-Ull [2021], Miquel-Vergés [2022] y Simó Soler [2023]) es una denominación genérica que se utiliza para designar contenido digital de vídeo, imágenes u otro completa o parcialmente fabricado, o la manipulación y transformación de contenido previamente existente (Van der Sloot & Wagenveld, 2022, p. 1) con la ayuda de aplicaciones que funcionan con base en técnicas de *deep learning* (Tolosana *et al.*, 2022, p. 4). Aunque la fabricación de imágenes o su manipulación son fenómenos antiguos, el uso de tecnología de IA ha permitido que el proceso de creación o manipulación se haga marcadamente más fácil, rápido y barato, facilitando obtener resultados más convincentes (Winter & Salter, 2020, p. 2). Estas técnicas han sido utilizadas en actividades como la sátira política, la generación de *fakenews* —informaciones falsas, algunas de ellas generadas con la ayuda de esas técnicas (Alibašić & Rose, 2019, p. 465; Melo, 2022)—, y de *pornfakes* —vídeos que muestran, generalmente, el rostro de mujeres famosas en escenas de contenido sexual o pornográfico (Harris, 2019; Brown & Fleming, 2020; Popova, 2020)—. A la fecha, el fenómeno de las *deepfakes* representa varios desafíos jurídicos y políticos (Chesney & Citron, 2019).

3 Los ML son modelizaciones de un idioma y de cómo este es utilizado por los seres humanos, estando compuesto por palabras con «significado» y reglas gramaticales, de sintaxis y de ortografía. La aplicación funciona con probabilidades, es decir, puede predecir qué palabra debería venir a continuación de otra, de acuerdo con un contexto determinado, gracias a *tokens* (unidades que corresponden a partes de ese idioma) y vectores de *embedding* (que permiten ir fijando vínculos entre las palabras). Tanto los *tokens* aprendidos como los requerimientos del usuario (*prompts*) son transformados a lenguaje de máquinas (secuencias numéricas binarias) que permiten a la aplicación reconocer tales secuencias y detectar patrones. Con base en esos patrones, el *chatbot* es capaz de generar texto que, al ser leído por un humano, es coherente y tiene lógica y sentido, aunque la información generada no sea necesariamente real o correcta. Dentro de los ML se distinguen los de «gran tamaño» (*Large Language Models*), concepto referido al volumen de datos usados en el proceso de entrenamiento de la aplicación y la cantidad de parámetros utilizados. Se puede encontrar una síntesis del tamaño de las bases de datos de entrenamiento de varias aplicaciones de ML en Bender *et al.* (2021, p. 611, tabla 1).

una «conversación», de ahí su descripción como «chat conversacional». En tal interacción, el usuario introduce un requerimiento (*script*) y obtiene una respuesta del chat, similar en forma y estructura a la que podría obtenerse de un interlocutor humano. ChatGPT es capaz no solo de incluir elementos de hecho y datos en sus respuestas, sino también de modificar sus respuestas con base en nuevos *inputs* proporcionados por el usuario. La amplia difusión de esta aplicación, su inmediata disponibilidad para uso por cualquier usuario con conexión a internet y la generación de ingentes volúmenes de investigación sobre ella justifican que en este artículo me refiera principalmente a ChatGPT. No obstante, los riesgos que se identifican son válidos, en general, para las tecnologías y aplicaciones de generación automatizada de contenido digital. Si se busca ampliar el tema, Gozalo-Brizuela y Garrido-Merchán (2023) hacen una revisión sobre el estado del arte en materia de tecnologías generativas.

Como generalmente ocurre con los avances tecnológicos, las aplicaciones generativas ofrecen innumerables ventajas⁴; pero, a la vez, imponen desafíos y representan riesgos que hacen imprescindible conocer y gestionar adecuadamente tales herramientas. El nacimiento de tecnologías generativas basadas en IA ha provocado alarmas en el mundo académico, prediciéndose el fin de la docencia como la conocemos hoy (Crawford *et al.*, 2023).

Este artículo se presenta como una primera aproximación a su objeto de análisis y abre una nueva línea de investigación de mi parte. Por esa razón, debe ser visto como una exposición descriptiva del fenómeno y, a la vez, un intento de revisión de la literatura sobre la materia. En ese contexto, tiene como objetivos describir los riesgos que representa para la integridad académica el desarrollo de tecnologías generativas basadas en IA y proponer algunas medidas de mitigación que se pueden adoptar para hacer frente a tales riesgos. En términos de la relevancia para este estudio, las tecnologías de generación de texto son capaces no solo de generar *abstracts* o sintetizar fuentes bibliográficas, sino también de escribir ensayos, artículos o tesis. Además, tienen la capacidad de responder preguntas incluidas en un examen, ya sea que estas estén expresadas en formato de selección múltiple o de desarrollo. Se ha sostenido que ChatGPT «puede generar, incluso en su estado básico, estudios de investigación de apariencia plausible para revistas bien ranqueadas» (Dowling & Lucey, 2023).

La enseñanza del derecho es especialmente vulnerable ante estas tecnologías. Los objetos de estudio de esta carrera se expresan en ideas abstractas que se fundamentan en datos de la realidad —como

4 Una revisión sistemática de estudios empíricos sobre el uso de ML de gran tamaño en aspectos educacionales puede encontrarse en Yan *et al.* (2023).

normas, evolución histórica de las instituciones, conflictos sociales, soluciones de aplicación de normas a hechos de la realidad, etc.—. Ese escenario es propicio para el uso (y abuso) de tecnologías generativas. En consecuencia, el foco de este artículo estará puesto en la docencia universitaria en la carrera de Derecho; sin embargo, ello no obsta a que los problemas que aquí se describen pueden ser extrapolables a otras áreas del conocimiento. Además, esta contribución puede servir como base para una —a mi juicio— necesaria discusión sobre estas temáticas que debe verificarse en las universidades.

En el apartado II expondré el contexto fáctico del objeto de estudio de este artículo, en el que haré una breve exposición sobre las aplicaciones de tecnologías algorítmicas de escritura y de generación de contenido, así como de los usos que puede dárseles en el ámbito del desempeño académico de estudiantes universitarios. El apartado III contiene una delimitación conceptual sobre la integridad académica con el objeto de precisar su contenido y, de esa forma, obtener elementos relevantes para el apartado IV. En este último hago una descripción de las conductas que se consideran disvaloradas desde la perspectiva de la integridad académica, distinguiendo entre plagio, suplantación de autoría y simulación de autoría. Asimismo, en el apartado V propongo una taxonomía de los riesgos que las tecnologías generativas representan para la integridad académica. Esta clasificación distingue entre riesgos procedimentales —excesiva dependencia y confianza en la AIG, irrealizabilidad del modelo pedagógico o formativo, e incorporación al aprendizaje de sesgos y discriminaciones— y riesgos de resultado —responsabilidad legal, aprendizaje distorsionado, sesgos y discriminación, sobrecalificación infundada, pérdida de competitividad de la institución de educación superior, contaminación de la cultura académica y degradación del valor ético de la institución—. En el apartado VI propongo cuatro medidas de mitigación de los riesgos previamente identificados en el apartado V —adaptación de los marcos normativos universitarios, revaloración del pensamiento crítico, cambios en los paradigmas de docencia y evaluación, y cuestionamiento de los entornos pedagógicos asincrónicos y a distancia—. El apartado VII, por último, está destinado a las conclusiones.

II. APLICACIONES DE GENERACIÓN AUTOMÁTICA DE LENGUAJE NATURAL

Las tecnologías de IAG, capaces de generar contenidos en forma autónoma y sin intervención humana directa, han registrado un avance importante en la interrelación directa con las personas. Es posible augurar que el lanzamiento de ChatGPT o Chat Generative Pre-Trained Transformer en noviembre de 2022 (Li *et al.*, 2023), desarrollado por la empresa OpenAI y considerado «un gran avance en la tecnología

235

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIALRISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

de IA generativa» (Eke, 2023)⁵, llegue a convertirse en un hito en esta materia. ChatGPT es la aplicación con el crecimiento de consumidores más acelerado de la historia⁶, con cien millones de usuarios registrados en dos meses desde su lanzamiento (Helberger & Diakopoulos, 2023)⁷. Además, ha logrado atraer la atención de legisladores⁸, medios de comunicación (Dale, 2021; García-Peñalvo, 2023), y académicos e investigadores, sobre todo en los aspectos éticos (Gordijn & Have, 2023). Recientemente, ChatGPT ha sido integrado con Bing, el motor de búsqueda de Microsoft, combinando las capacidades conversacionales con las de búsqueda en internet (Bing Chat); y se ha liberado una nueva versión: ChatGPT-4⁹. ChatGPT no es el único chat conversacional (Floridi, 2023, p. 14)¹⁰, pero su lanzamiento aceleró una carrera entre las empresas tecnológicas en la materia (Höppner & Streatfeild, 2023; Rahaman *et al.*, 2023)¹¹.

ChatGPT es una herramienta capaz de procesar imágenes y lenguaje humano y de producir contenido. Es una interfaz conversacional en línea basada en un paradigma de modelo de generación de lenguaje preentrenado (Susnjak, 2022, p. 3; Kasneci *et al.*, 2023)¹². «Está diseñado para generar respuestas de texto similares a las humanas a la entrada del usuario en un contexto conversacional» (Aydin & Karaarslan, 2022, p. 22), con niveles crecientes de confiabilidad a medida que se

5 En junio de 2018, científicos de OpenAI publicaron el *paper* «Improving Language Understanding by Generative Pre-Training» en versión *preprint* en la propia web de OpenAI (Radford *et al.*, 2018), por entonces, una organización sin fines de lucro. El trabajo, firmado por Alec Radford, Karthik Narasimhan, Tim Salimans e Ilya Sutskever, es el comienzo de ChatGPT.

6 Ver Bartz (2023).

7 Ver Milmo (2023a).

8 Helberger y Diakopoulos (2023) analizan los ajustes que la Unión Europea está intentando introducir a su proyecto de regulación sobre IA a partir de la difusión de ChatGPT y cómo esta tecnología no calza en los moldes tradicionales con los que se analiza y regula la IA en general.

9 La versión ChatGPT-4 fue anunciada el 15 de marzo de 2023. Ella es descrita por OpenAI como un «gran modelo multimodal (que acepta entradas de texto y emite salidas de texto hoy, con entradas de imágenes en el futuro) que puede resolver problemas difíciles con mayor precisión que cualquiera de nuestros modelos anteriores, gracias a su conocimiento general más amplio y capacidades de razonamiento avanzadas» (OpenAI, 2023). Sin embargo, la compañía ha sido reuente en proporcionar públicamente información que respalde las anunciadas características de su nuevo producto (Heaven, 2023).

10 Entre los otros grandes modelos de lenguaje se pueden mencionar BERT, GLaM y LaMDA, todos de Google; Fairseq y LLaMA, de Meta; GPT-Neo, GPT-J y GPT-NeoX, de EleutherAI; Claude, de Anthropic; Megatron-Turing NLG, de Microsoft y Nvidia; Chinchilla, de DeepMind; YaLM 100B, de Yandex; BLOOM, de Hugging Face; y Alexa, de Amazon.

11 Google, por ejemplo, ha tenido varios problemas con su *chatbot* conversacional Bard y la tecnología que lo soporta, LaMDA (*Language Model for Dialogue Applications*). Además de los errores en los que incurrió sobre el telescopio espacial James Webb en su presentación pública (Milmo, 2023b), que le significó a Google una fuerte baja del precio de sus acciones (Floridi, 2023, p. 15), la firma tecnológica debió enfrentar el conflicto generado por uno de sus ingenieros, que decidió hacer públicas sus preocupaciones por la eventual capacidad sintiente que habría desarrollado el *chatbot* (Tiku, 2022). Por su parte, Meta, propietaria de Facebook, Instagram y WhatsApp, ha efectuado anuncios sobre su propia aplicación conversacional basada en *deep learning*, que podría llamarse LLaMa (Herr, 2023).

12 Este modelo de generación de lenguaje preentrenado se basa en la arquitectura *Transformer*, es el de más reciente desarrollo y muestra mejores rendimientos que los otros tres modelos existentes —el basado en ejemplos, el estadístico y el basado en redes neuronales— en la ejecución de tareas. Su mayor debilidad se encuentra, sin embargo, en que tiene menores niveles de fidelidad que los otros modelos (Li *et al.*, 2022).

va profundizando su entrenamiento directamente proporcionales a la generalidad de la materia sobre la que se le consulte (Qin *et al.*, 2023). ChatGPT, de uso gratuito¹³ a través de internet¹⁴, funciona gracias a algoritmos de *deep learning*, un tipo de IA. Tiene la habilidad de «comprender» las preguntas o los requerimientos que se le formulen, encontrar patrones o relaciones generadas a partir de las enormes bases de datos con las que fue entrenado —y sigue siéndolo—¹⁵, y construir una respuesta escrita con diferentes niveles de corrección y precisión.

Desde la aparición pública de ChatGPT se han hecho estudios y ensayos que pretenden analizar la posibilidad de que dicha herramienta pueda generar textos académicos. Susnjak (2022) experimentó con diversas respuestas generadas por ChatGPT. Al analizarlas a la luz de los estándares intelectuales propuestos por Paul (2005), concluyó que las respuestas proporcionadas por el chat «pueden evaluarse como claras en la exposición, precisas con respecto a los ejemplos utilizados, relevantes para las preguntas formuladas y suficientemente profundas y amplias considerando las restricciones impuestas y lógicamente coherentes en textos más extensos» (p. 13). Jabotinsky y Sarel (2023) condujeron un experimento en el que requirieron a ChatGPT que generara texto, incluyendo referencias bibliográficas, sobre los riesgos de la IA en la producción científica. Los autores lograron determinar que el *chatbot* fue capaz de generar textos con sentido y bien escritos. Sin embargo, detectaron errores en las referencias bibliográficas proporcionadas (algunas de ellas, inexistentes) y destacaron que la información con la que ha sido entrenado solo comprende hasta 2021. A pesar de los defectos, los autores concluyen que ChatGPT tiene un potencial útil para la escritura académica. Asimismo, los autores sometieron a prueba a Bing y encontraron que este *chatbot* generó información menos específica y utilizó fuentes de información generales (como blogs) que no califican como fuentes válidas de información para un *paper* académico.

Bommarito y Katz (2022) condujeron una evaluación experimental de ChatGPT-3 (la versión previa a ChatGPT) que consistió en aplicarle una prueba de acceso a la barra de abogados en Estados Unidos. Los autores descubrieron que el rendimiento de ChatGPT en resolver esos exámenes se ubicó muy por encima del margen de aleatoriedad en los resultados y que superó el porcentaje de aprobación en dos de las materias incluidas en la evaluación (derecho de daños y de evidencia).

13 Aunque existe la posibilidad de pagar una suscripción que da acceso a la versión Plus con mejoras en rendimiento y tiempos de respuesta respecto de la versión gratuita. Se puede a ceder a ChatGPT a través del siguiente enlace: <https://openai.com>

14 Ya sea desde un navegador o *browser* en un computador de escritorio, un dispositivo móvil (como un *smartphone* o una *tablet*), la web de OpenAI o a través de API y extensiones desarrolladas para integrar ChatGPT con otras aplicaciones.

15 Se estima que ChatGPT fue entrenado con un conjunto de datos compuesto por trescientos mil millones de palabras (Helberger & Diakopoulos, 2023).

También hallaron que tuvo un rendimiento muy similar al de postulantes humanos. Carrasco *et al.* (2023) sometieron a ChatGPT al examen MIR 2022 en España y concluyeron que, de acuerdo con el puntaje obtenido en tal examinación, el *chatbot* de OpenAI podría «haber elegido múltiples especialidades en distintos hospitales a lo largo del Estado» español. Choi *et al.* (2023), haciendo un ensayo similar, determinaron que ChatGPT habría aprobado los exámenes finales de las cuatro asignaturas que incluyeron en su estudio, obteniendo una calificación C+ en todas ellas; y que «si dicho desempeño fuera constante en toda la carrera de derecho, las calificaciones obtenidas por ChatGPT serían suficientes para que se graduara».

En el proceso de redacción de este artículo, desarrollé un pequeño ensayo solo para ilustrar el rendimiento de ChatGPT y sin pretender que sus resultados sirvan como evidencia porque no apliqué una metodología específica. Le formulé a ChatGPT sesenta de las preguntas del Modelo de Prueba PDT de acceso a las universidades chilenas para las asignaturas de Historia y Ciencias Sociales publicado por Demre, tal como aparecen escritas en el modelo. La aplicación reconoció el formato de selección múltiple de las preguntas, ya que en sus respuestas indicó cuál de las alternativas era la correcta; pudo completar cuadros de información (como en la pregunta N.º 27 del modelo); fue capaz de interpretar diagramas presentados en el modelo gráficamente, pero introducidos a ChatGPT solo con texto (como en las preguntas N.º 34, 40, 49 y 51); realizó ejercicios de comprensión lectora (preguntas N.º 9, 23, 29, 30, 32, 44, 47, 48, 50 y 55); y pudo prescindir de gráficos de información incluidos en el modelo de prueba, pero indicando la fuente de los datos (preguntas N.º 20 y 58). Al comparar las respuestas entregadas por ChatGPT¹⁶ con las claves publicadas por Demre (2022), ChatGPT acertó en cincuenta y cinco de las sesenta preguntas formuladas; esto es, tuvo un 91,7 % de acierto. Esto significa que, de haber rendido la examinación, ChatGPT habría obtenido ochocientos ochenta y ocho puntos de un total de mil.

En el mundo del derecho, y a pesar de su enfoque generalista, ChatGPT ya ha tenido usos concretos en la función jurisdiccional. En esa línea, este *chatbot* fue utilizado por un tribunal colombiano como apoyo en la construcción de fundamentos jurídicos en el dictado de una de sus sentencias¹⁷.

¹⁶ Generadas con la versión liberada el 14 de marzo de 2023 (OpenAI, s.f.).

¹⁷ Se trata de la sentencia del Juzgado 1.º Laboral del Circuito de Cartagena, de 30 de enero de 2023. En el proceso se discutía el alcance de las prestaciones de seguridad social para un niño con síndrome autista. Recurriendo a la Ley 2213 del 13 de junio de 2022, «Por medio de la cual se establece la vigencia permanente del decreto legislativo 806 de 2020 y se adoptan medidas para implementar las tecnologías de la información y las comunicaciones en las actuaciones judiciales, agilizar los procesos judiciales y flexibilizar la atención a los usuarios del servicio de justicia y se dictan otras disposiciones», el Tribunal —luego de expresar sus fundamentos jurídicos—, agrega que se decidió «extender los argumentos de la decisión adoptada, conforme al uso de inteligencia artificial (IA)».

Floridi (2023) explica los ML comparándolos por analogía con los loros estocásticos, pero advirtiendo brillantemente que

La analogía ayuda, pero solo en parte, no solo porque los loros tienen una inteligencia propia que sería la envidia de cualquier IA sino, sobre todo, porque los [Modelos de Lenguaje de Gran Tamaño] sintetizan textos de nuevas formas, reestructurando los contenidos en los que han sido entrenados, no proporcionando simples repeticiones o yuxtaposiciones. Se parecen mucho más a la función de autocompletar de un motor de búsqueda. Y en su capacidad de síntesis se asemejan a esos alumnos mediocres o perezosos que, para escribir un breve ensayo, utilizan una decena de referencias pertinentes sugeridas por el profesor y, tomando un poco de aquí y un poco de allá, arman un texto ecléctico, coherente, pero sin haber entendido mucho ni añadido nada. Como tutor universitario en Oxford, corregí muchos de ellos cada trimestre. Ahora ChatGPT puede producirlos de manera más rápida y eficiente (p. 15).

Estas tecnologías ofrecen oportunidades únicas para el mejoramiento continuo de la enseñanza universitaria que «incluyen una mayor participación, colaboración y accesibilidad de los estudiantes» (Cotton *et al.*, 2023). Pueden permitirles identificar fuentes de información de una manera más rápida que a través de las búsquedas en bases de datos convencionales, facilitarles el procesamiento de mayores volúmenes de datos para discriminar los relevantes de los accesorios o contribuir al proceso de formación de un juicio crítico sobre diversos temas, por mencionar algunos aspectos. Pero estas tecnologías también plantean desafíos (Islam & Islam, 2023; Kasneci *et al.*, 2023; Riley & Alvarez, 2023), entre ellos uno de carácter ético e, incluso, jurídico que se relaciona con lo que en términos coloquiales denominamos «copia».

La versión 3.5 de ChatGPT, disponible en forma gratuita, se diferencia de todas las otras aplicaciones generativas y conversacionales existentes, incluso de sus propias versiones previas¹⁸, por las capacidades que dicho *chatbot* tiene. No solo puede entender texto natural y contestar

239

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIALRISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

La sentencia deja constancia de que se le formularon varias preguntas a ChatGPT sobre la cuestión debatida como, por ejemplo: «¿El menor autista está exonerado de pagar cuotas moderadoras en sus terapias?», «¿Las acciones de tutela en estos casos se deben conceder?», «¿Exigir en estos casos la cuota moderadora es una barrera de acceso al servicio de salud?» y «¿La jurisprudencia de la Corte Constitucional ha tomado decisiones favorables en casos similares?». La sentencia consigna las respuestas generadas por la aplicación y advierte que «[e]l propósito de incluir estos textos producto de la IA no es en manera alguna reemplazar la decisión del Juez. Lo que realmente buscamos es optimizar los tiempos empleados en redacción de sentencias, previa corroboración de la información suministrada por IA» (Sentencia de 30 de enero de 2023, §11). En todo caso, no se alcanza a visualizar de qué forma el uso de ChatGPT podría lograr los propósitos declarados, sobre todo si en el proceso de dictado de la sentencia el Tribunal recurre a la aplicación solo después de exponer los fundamentos construidos por el juez. De todos modos, este uso concreto de la IA fue objeto de atención mediática; por ejemplo, el hecho fue cubierto por *The Guardian* (Taylor, 2023).

18 Generative Pre-trained Transformer 2 (GPT-2) de OpenAI, lanzado en noviembre de 2019, y el Generative Pre-trained Transformer 3 (GPT-3), de la misma empresa y lanzado en junio de 2020 (Zhang & Li, 2021; Perkins, 2023).

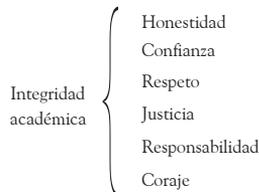
preguntas, sino que también es capaz de traducir entre distintos idiomas, rechazar requerimientos inapropiados, contradecir premisas incorrectas, resumir textos, generar escritura creativa como poesía, contar chistes o crear cuentos (Eke, 2023; Jiao *et al.*, 2023; Talan & Kalinkara, 2023, p. 35). Y se encuentra aún en su infancia. Como lo han destacado Aljanabi y ChatGPT (2023, p. 16) —isí, el propio *chatbot* es coautor declarado de esta fuente!—, «una de las direcciones futuras más prometedoras para ChatGPT es su integración con otras tecnologías de IA, tales como la visión de computadores y la robótica»¹⁹. A ello se puede agregar que ChatGPT tiene un enfoque generalista con relación a la información de entrenamiento empleada, pero nada obsta a que en un corto o mediano plazo surjan chats conversacionales entrenados en áreas específicas del conocimiento. El derecho puede ser una de ellas.

III. INTEGRIDAD ACADÉMICA

Para identificar los riesgos que las tecnologías de IA de generación de contenido representan para la integridad académica²⁰, tomaré como referencia la definición propuesta por el International Center for Academic Integrity (2021) sin perjuicio de destacar que, a nivel de definición, un consenso sobre cómo definir la integridad académica está lejos de alcanzarse y su comprensión está fuertemente influida por contextos nacionales, como lo demuestran los diferentes trabajos que integran el *Handbook of Academic Integrity* (Bretag (ed.), 2016).

El International Center for Academic Integrity (2021) define la integridad académica como un compromiso con seis valores fundamentales: honestidad, confianza, justicia, respeto, responsabilidad y coraje (ver figura 1).

Figura 1. Esquema de los valores que integran el concepto de integridad académica



Fuente: International Center for Academic Integrity (2021).

¹⁹ Como las funcionalidades que ya ha desarrollado ChatGPT-4. Al respecto, ver nota al pie N.º 9.

²⁰ García-Villegas *et al.* (2016) explican que en el contexto latinoamericano se prefieren otras expresiones para dar cuenta del fenómeno, tales como «fraude» o «incumplimiento de deberes académicos»; es decir, denominaciones que ponen el acento en sus aspectos negativos a diferencia de la de «integridad académica», más propia del entorno angloamericano, que destaca un enfoque positivo (pp. 162-163).

Desde una perspectiva deontológica, la adhesión sustancial a estos valores permite reconocer una actividad académica como tal en la medida que la adecuación a tales valores es un requisito de la esencia de las labores de docencia e investigación. Desde un enfoque ético utilitarista, la recepción de los señalados valores permite atribuirle valor y credibilidad a la actividad académica.

El International Center for Academic Integrity (2021) estima que el valor de la honestidad se encuentra en la base de los valores de la confianza, la justicia, el respeto y la responsabilidad, actuando como un sustrato necesario para el desarrollo ético. Entiende a la honestidad como aquella cualidad positiva que consiste en ser legítimo o veraz, esto es, como la ausencia de fraude o engaño.

En virtud del valor de la confianza, los «miembros de la comunidad académica deben ser capaces de confiar en que el trabajo, ya sea de estudiantes o investigadores, no es falsificado y que los estándares se aplican por igual a todos» (p. 6).

El valor de justicia es definido por ICAI como aquella cualidad positiva que consiste en la «calidad o estado de ser justo, especialmente el trato justo o imparcial, la falta de favoritismo hacia un lado u otro» (p. 7). La adhesión a este valor impone el deber de un tratamiento imparcial que refuerza la importancia de la verdad, las ideas, la lógica y la racionalidad; y se manifiesta en componentes de predictibilidad, transparencia y expectativas claras y razonables (p. 7).

El valor del respeto es entendido como una alta o especial consideración o la cualidad de ser estimado. Este valor se desenvuelve en dos perspectivas. La primera, el autorrespeto, que significa enfrentar los desafíos sin comprometer los valores propios; y la segunda, el heterorrespeto, que consiste en «valorar la diversidad de opiniones y apreciar la necesidad de desafiar, probar y refinar las ideas» (p. 8).

El valor de la responsabilidad consiste en la cualidad de asumir un deber y cumplirlo desde una perspectiva ética e, incluso, normativa (jurídica). La adhesión del valor de la responsabilidad para con uno mismo como para con los otros distribuye y aumenta la posibilidad de introducir cambios importantes en la comunidad académica (p. 9).

Finalmente, el valor del coraje significa la capacidad de cada uno de los miembros de la comunidad de actuar de acuerdo con los propios valores con fundamento, a pesar del miedo de adoptar esa posición y de comportarse de acuerdo con las propias convicciones personales. Tal actuación solo tiene protección si los valores propios tienen fundamento y son aceptados por el resto de la comunidad (p. 10).

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIAL

RISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

IV. CONDUCTAS Y SU (DIS)VALORACIÓN

Desde la aparición de los programas de procesamiento de texto hemos contado con herramientas de ayuda en el proceso de escritura; por ejemplo, las funciones de corrección ortográfica y gramatical y de autocorrección con que cuenta el procesador de texto MS Word, o los asistentes digitales de escritura (DWA, por sus siglas en inglés), que permiten introducir mejoras sustanciales a la redacción, la gramática y el vocabulario utilizado en un texto escrito. El anuncio efectuado por Microsoft sobre el lanzamiento de 365 Copilot²¹ promete llevar las herramientas productivas a un nivel sin precedentes.

Las utilidades existentes son herramientas cuyo uso no solo se encuentra naturalizado, sino que, además, puede ser valorado. Nadie cuestionaría a un estudiante que ha enmendado sus erratas en un ensayo o en una tesis gracias al corrector ortográfico y gramatical de MS Word; y, de hecho, lo reprochable sería que, en la redacción de un trabajo escrito, su autor incurra en tales errores por no haberlas utilizado. Los asistentes digitales de escritura basados en IA ayudan a superar barreras idiomáticas que pueden incrementar el acceso de personas no nativas del inglés para publicar sus investigaciones en revistas anglófonas que, prácticamente, monopolizan las principales vías de difusión del conocimiento, incluso en las ciencias jurídicas y criminológicas (Faraldo-Cabana, 2018; 2019; Faraldo-Cabana & Lamela, 2021). Las aplicaciones de ML, a través de las funciones para las que están diseñadas, tienen un potencial enorme, cuya utilización en la docencia podría aportar gran valor.

No obstante, las herramientas informáticas también generan riesgos para la integridad académica y al propio proceso docente que es necesario identificar y mitigar. Tal tarea debe comenzar por una identificación previa respecto de las posibles conductas involucradas. No cualquier utilización de herramientas basadas en IA va a conducir, automáticamente, a conductas que pueden desvalorarse desde la perspectiva de la integridad académica. Establecer dicha relación causal sería una simplificación inadmisibile.

En palabras de Eke (2023), el riesgo de las aplicaciones ML consiste en que «tanto los estudiantes como los investigadores puedan comenzar a externalizar su escritura a ChatGPT. Si algunas respuestas tempranas a las preguntas de ensayo de nivel universitario sirven de algo, los profesores deberían estar preocupados por el futuro de los ensayos como una forma de evaluación».

²¹ Microsoft (2023) ha anunciado la integración de IA a su paquete de aplicaciones de productividad 365 Copilot, antes conocida como Office. Esto significa que Word, Excel, PowerPoint, Outlook y Teams, entre otras, pasarán a estar dotadas de la capacidad de desarrollar acciones en forma autónoma.

Me parece acertada la descripción que hace Eke porque permite distinguir dos grandes grupos de casos dentro de la utilización de tecnologías. El primero, compuesto por aquellas conductas en las que el uso de herramientas tecnológicas es irreprochable porque tal utilización no lesiona valores ni reglas; el segundo, constituido por comportamientos desvalorados porque el uso de herramientas tecnológicas infringe tales valores o normas. En este grupo, que Dawson (2021) denomina genéricamente *e-cheating*, el uso de las herramientas tecnológicas sobrepasa el mero apoyo o ayuda que ellas prestan. Estimo que en este grupo de *e-cheating* pueden reconocerse tres formas específicas de conductas desvaloradas: plagio, suplantación de autoría y simulación de autoría.

En el caso del plagio —palabra que derivaría del latín *plagiare*, «secuestro con el propósito de convertir a la víctima en esclavo» (Fishman, 2016, p. 13)—, es poco lo que puede agregarse en términos de su descripción como conducta desvalorada, a pesar de que en ella no abundan los consensos ni las definiciones precisas (Jelenic & Kennette, 2022, pp. 16-17). Según una sentencia penal dictada en Chile, el plagio es «[l]a apropiación de una obra ajena, haciéndola pasar como propia o bien utilizando los elementos creativos de aquella para la elaboración de una nueva creación de carácter ilegítima» (*Carlos Antonio Urquieta Salazar c/ Margarita Leonor Cid Lizondo*, 27 de mayo de 2011, considerando 11). De acuerdo con la Organización Mundial de Propiedad Intelectual (OMPI, 1980), el plagio consiste en «el acto de ofrecer o presentar como propia, en su totalidad o en parte, la obra de otra persona, en una forma o contexto más o menos alterados» (p. 192). El elemento central del plagio es la usurpación de un contenido perteneciente a otra persona que es su autora.

El plagio convencional o básico consiste en la transcripción de un contenido de autoría ajena y su atribución a una persona distinta. Aunque las tecnologías de IA pueden facilitar esta forma de plagio, es evidente que la posibilidad de incurrir en él es completamente independiente de tales tecnologías. En otras palabras, el riesgo de plagio convencional existe con o sin tecnologías de IA e, incluso, con la misma intensidad. Con relación a las herramientas informáticas, el riesgo de plagio es especialmente relevante a propósito de las herramientas de paráfraseo automático (APT, por sus siglas en inglés). El plagio puede ser facilitado sustancialmente por estas herramientas (Rogerson & McCarthy, 2017), ya que ellas permiten cambiar la estructura sintáctica y gramatical de párrafos enteros, pero adoptando las ideas contenidas en la fuente paráfraseada y omitiendo la referencia a la fuente consultada. El paráfraseo no es necesariamente una tarea fácil y requiere no solo de habilidades, sino también de conocimientos sobre la materia; pero las tecnologías de IA permiten a quien efectúa el paráfraseo prescindir de

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIAL

RISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

habilidades y conocimientos para hacerlo. Esta modalidad, denominada «plagio simulado», «inteligente» o «elaborado» (Echavarría, 2014, p. 710; 2016), puede burlar fácilmente los *softwares* de detección de plagio.

En el entorno angloamericano se ha investigado y descrito la conducta de *commercial cheating*, esto es, el reclutamiento por parte de estudiantes de terceras personas que escriben por ellos o ellas sus ensayos, memorias o tesis a cambio de un pago (Newton & Lang, 2016; Newton, 2018) y que en nuestro medio se conoce como «suplantación de autoría» (Comas Forgas & Sureda Negre, 2008; Rojas Chavarro & Olarte Collazos, 2010)²². No conozco estudios que hayan determinado la incidencia de este fenómeno en Chile, pero su existencia es una cuestión conocida²³. La externalización de la redacción de trabajos académicos a una aplicación de generación automática de texto funciona con la misma lógica del *commercial cheating* y constituye el mismo tipo de infracción a la integridad académica. No obstante, las aplicaciones de generación automática de texto pueden llevar el fenómeno a niveles sin precedentes y tienen el potencial de amenazar masivamente la forma en la que entendemos la docencia y el proceso de evaluación. En palabras de King y del propio ChatGPT (2023),

[L]os estudiantes universitarios podrían potencialmente usar ChatGPT para hacer trampa en las tareas de redacción de ensayos alimentando al chatbot con indicaciones y preguntas específicas, y luego copiando y pegando las respuestas generadas en sus propios trabajos. Esto les permitiría producir fácilmente ensayos sin poner el esfuerzo y la investigación necesarios para escribir contenido original.

No obstante el plagio y la suplantación de identidad, las tecnologías de IA analizadas en este artículo plantean un escenario diferente. En términos descriptivos, este nuevo escenario consiste en el caso de un estudiante que haya obtenido su trabajo directamente desde una aplicación de generación automática de texto, o una parte más o menos sustancial de este, pero que aparente ser su autor, sin reconocer su verdadera fuente de creación, sin referenciar el origen de los insumos para su redacción o sin declarar el método utilizado para la creación del contenido.

En este escenario, la extracción sin indicar la fuente no puede considerarse plagio porque el *chatbot* no es un ser humano y el plagio

22 El problema parece presentarse no solo a nivel de estudios conducentes a un grado, sino también en los estudios de doctorado. Aitchison y Mowbray (2016) han estudiado «servicios de apoyo a la escritura», generalmente provistos *online* y especialmente dirigidos a estudiantes de doctorado que sucumben ante la constante presión por publicar.

23 Basta una simple búsqueda en internet para encontrar sitios en los que se ofrece el servicio de «asesoramiento» en la redacción de memorias y tesis. Es muy probable que la fórmula de la asesoría sea un eufemismo para referirse a la venta de textos escritos por terceras personas que el adquirente presentará como elaborados por él mismo.

requiere de la usurpación de la autoría de una persona (Kleebayoon & Wiwanitkit, 2023). Es cierto que el *chatbot* habrá tomado la información de alguna fuente bibliográfica humana, ya que su habilidad no es inventar una respuesta o generar nuevo conocimiento, sino articular una respuesta sobre la base de su información de entrenamiento. Pero, aun así, tomar la información preparada por un *chatbot* que proviene de fuentes «humanas» y simular ser su autor no es suficiente para considerarlo una forma de plagio propiamente tal. Asimismo, en este escenario no hay una contraprestación para el tercero que actúa como el verdadero autor del contenido; por el contrario, solo existe una relación entre el estudiante y una máquina inteligente generadora del contenido. Propongo denominar «simulación de autoría» a esta modalidad de deshonestidad académica que se desarrolla en entornos tecnológicos.

A modo de ilustración, formulé a ChatGPT el siguiente requerimiento: «Escribe un resumen sobre “Who Dares Fine a Murderer? The Changing Meaning of Money and Fines in Western European Criminal Systems”, escrito por Patricia Faraldo-Cabana» (Faraldo-Cabana, 2015). El *chatbot* generó la siguiente respuesta:

“Who Dares Fine a Murderer? The Changing Meaning of Money and Fines in Western European Criminal Systems” es un artículo escrito por Patricia Faraldo-Cabana que aborda el papel del dinero y las multas en los sistemas criminales de Europa occidental. El artículo examina cómo ha cambiado la percepción del dinero como una forma de compensación por el delito, desde la antigua Grecia hasta la actualidad, y cómo el uso de multas ha evolucionado en respuesta a los cambios económicos y culturales.

Faraldo-Cabana argumenta que la multa se ha utilizado históricamente como una forma de restaurar el orden social, ya que la compensación monetaria no solo compensaba a la víctima sino que también castigaba al delincuente. Sin embargo, a medida que la economía se ha vuelto más compleja, la multa se ha utilizado cada vez más como una fuente de ingresos para el Estado en lugar de un castigo por el delito.

El artículo también destaca las diferencias culturales en la percepción del dinero y las multas en Europa occidental, y cómo estas diferencias han influido en la evolución de los sistemas criminales. Por ejemplo, en países como Alemania, donde se valora la idea de que la justicia no tiene precio, las multas se han utilizado con moderación, mientras que en países como España, donde las multas se han utilizado históricamente como una forma de obtener ingresos para el Estado, se han aplicado con mayor frecuencia.

En general, “Who Dares Fine a Murderer? The Changing Meaning of Money and Fines in Western European Criminal Systems” ofrece una perspectiva interesante sobre cómo la percepción del dinero y las

245

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIALRISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

multas ha evolucionado en los sistemas criminales de Europa occidental y cómo estos cambios han reflejado las transformaciones económicas y culturales a lo largo del tiempo²⁴.

El texto puede considerarse una buena síntesis del artículo, pero ChatGPT erró en la identificación de la fuente²⁵. Sin embargo, su respuesta perfectamente podría ser utilizada por un o una estudiante en la entrega de un informe o trabajo escrito evaluado.

Surge la interrogante sobre la necesidad de distinguir una forma específica de conductas disvaloradas —la simulación de autoría— si ya existen el plagio y la usurpación de autoría. Estimo que reconocer esta tercera modalidad de conducta disvalorada se justifica por las diferencias que tiene con relación a las otras dos modalidades.

La primera diferencia consiste en que las tecnologías de IA tienen la capacidad de conferirle a la simulación de autoría un efecto amplificador de los riesgos para la integridad académica. Si el plagio tradicional se basa en la fatídica, pero engorrosa combinación de las acciones copiar-pegar (*copy-paste*), las tecnologías algorítmicas de escritura y de generación de contenido que se pueden usar deshonestamente pueden serlo a una escala industrial por la facilidad de uso que ellas tienen y la posibilidad de replicar sus utilizaciones. Haciendo una analogía, el aumento de la escala de los riesgos para la integridad académica es igual al que, en su momento, representó la invención y masificación de la máquina fotocopiadora para la propiedad intelectual de los libros.

La segunda diferencia que plantean las indicadas tecnologías es que su utilización es prácticamente indetectable y que, con ello, resulte imposible acreditar una conducta disvalorada. La indetectabilidad se debe a que los *softwares* de detección de plagio no son capaces de identificar el parafraseo hecho con herramientas algorítmicas, ni tampoco si un texto fue producido por una aplicación de generación automatizada (Rudolph *et al.*, 2023). El profesor o profesora que evalúa un trabajo escrito tampoco está en condiciones de percibir si un texto es de autoría propia del estudiante sometido a la evaluación o fue escrito por una máquina. Se han hecho esfuerzos por desarrollar herramientas de detección (clasificación) de contenido generado automáticamente, pero sin resultados concretos por el momento (Perkins, 2023, p. 13). OpenAI, de hecho, informa en su página web que se encuentra trabajando en una aplicación denominada Classifier (Kirchner *et al.*, 2023), que podría llegar a tener la capacidad de determinar si un texto fue escrito por una persona o fue generado por ChatGPT. Pero la empresa advierte que

24 Texto generado por ChatGPT Mar 14 version. Se puede revisar la aplicación en el siguiente enlace: <https://chat.openai.com/>

25 Según la aplicación, el artículo fue publicado en *Law and Society Review* en 2017 (la referencia correcta puede consultarse en la bibliografía).

dicha aplicación tiene importantes limitaciones²⁶ y que ella «[n]o debe usarse como una herramienta principal para la toma de decisiones, sino como un complemento de otros métodos para determinar la fuente de un texto» (Kirchner *et al.*, 2023). Además, advierten los ingenieros de OpenAI, «los estudiantes pueden aprender rápidamente cómo evadir la detección modificando algunas palabras o cláusulas en el contenido generado» (Kirchner *et al.*, 2023). La indetectabilidad también se debe a los altos costes asociados a los procesos de auditoría o investigación sobre la autoría y detección de conductas disvaloradas, que requieren de tecnologías innovadoras que no todas las universidades están en condiciones de financiar o en capacidad de acceder a ellas (Crawford *et al.*, 2023, p. 2; Eke, 2023, p. 3; Kasneci *et al.*, 2023, p. 13).

La característica de la indetectabilidad y la precariedad de las herramientas para contrarrestarla hacen que surja otro tipo de riesgo: la posibilidad de juzgar inadecuadamente como conducta disvalorada el comportamiento de un estudiante que no lo ha sido (falso positivo) o la de considerar como lícito el comportamiento de un estudiante que efectivamente ha sido deshonesto (falso negativo). Las tecnologías generativas pueden desencadenar un círculo vicioso de afectación al valor de la confianza como parte de la integridad académica. Por un lado, la posibilidad de su utilización por los y las estudiantes podría conducir a los profesores y las profesoras a comenzar el proceso evaluativo desde una premisa de desconfianza; es decir, desde una situación inicial de negación de dicho valor. Por otro lado, la urgencia por detectar y evitar conductas disvaloradas plantea el riesgo de instalar culturas de vigilantismo al interior del proceso docente que también implican una afectación al valor de la confianza como elemento estructurante de la integridad académica²⁷.

La tercera diferencia consiste en el carácter informatizado de las técnicas de deshonestidad que el mal uso de las tecnologías de IA permite desarrollar (Dawson, 2021, pp. 5-6). Esto significa que, en tanto los métodos «tradicionales» responden a experiencias individuales difícilmente compartibles con otros en términos de sus productos, las tecnologías están ampliamente disponibles, son transables, distribuibles y accesibles para casi cualquier usuario que disponga de una conexión a internet y un dispositivo que le permita navegar por ella. En otras palabras, las técnicas de deshonestidad tecnologizadas dejan de ser una experiencia individual y pueden pasar a tener un carácter colectivo que llegue a muchas personas instantáneamente.

26 Por ejemplo, es muy poco fiable en textos cortos de menos de mil caracteres, pues puede generar falsos positivos o negativos. Además, se recomienda su uso solo para textos escritos en idioma inglés, ya que en otros idiomas «funciona significativamente peor» (Kirchner *et al.*, 2023). Tampoco puede operar sobre textos predecibles, como una lista de los primeros mil números primos, porque la respuesta siempre la misma.

27 Este riesgo ha sido puesto de relieve por Henry y Oliver (2022).

V. RIESGOS

La identificación de los riesgos éticos, su evaluación y mitigación son áreas que han adquirido especial relevancia a propósito de la energía que ha mostrado el desarrollo de la IA después de su segundo invierno (Boddington, 2017; Bryson, 2020; Donath, 2020; Eitel-Porter, 2021; Felländer *et al.*, 2022; Hagendorff, 2020; 2022; Hermansson & Hansson, 2007; Powers & Ganascia, 2020; Siau & Wang, 2020; Spivak & Shepherd, 2021; Vanem, 2012; Yapo & Weiss, 2018). En general, se destacan los riesgos que plantea la IA para la vigencia de los derechos fundamentales (Council of Europe, 2018; Yeung, 2019), o para la decisión y ejecución de políticas públicas que podrían imponer una *algorocracia* (Danaher, 2016; 2022; Chace, 2018, pp. 89 y ss.; Gritsenko & Wood, 2020; Chomanski, 2022), entre otros tópicos. En particular, los riesgos para la integridad académica también han despertado la atención científica y académica (Abdelaal *et al.*, 2019; Jabotinsky & Sarel, 2023; Lancaster, 2021; Moya *et al.*, 2023; Perkins, 2023; Roe & Perkins, 2022).

Los riesgos deben ser entendidos como amenazas que pueden llegar a concretarse en daños o afectaciones efectivas, en el caso del objeto de este artículo, para la integridad académica. Es decir que la IA tiene ciertas condiciones que, en determinadas circunstancias, representan una potencial agresión a los valores sobre los que se estructura la integridad académica.

La identificación de los riesgos de la IA en la docencia superior requiere reconocer que ella tiene una doble dimensión. Por un lado, es un proceso continuo de carácter dinámico que se va desarrollando en el tiempo y en el que se van verificando hitos; y, por otro, es un resultado, una situación final a la que se llega. Por ello, he identificado dos grandes grupos de riesgos: los *procedimentales*, que refieren a las amenazas sobre la docencia entendida como proceso; y los *de resultado*, es decir, los peligros de la IA que impactan el producto final del proceso de docencia²⁸. Un esquema de los riesgos identificados se muestra en la figura 2.

28 Se puede encontrar una tipología diferente de riesgos para la integridad académica en Kasneci *et al.* (2023).

Figura 2. Esquema de los riesgos de la IA para la integridad académica



Fuente: elaboración propia.

No es posible en esta oportunidad hacer una descripción detallada de cada uno de los riesgos identificados porque dicha empresa excedería la extensión de un artículo. Por ello, formularé una enunciación de tales riesgos, dejando la posibilidad de su desarrollo para futuras investigaciones.

V.1. Riesgos procedimentales

V.1.1. Excesiva dependencia y confianza en la AIG

La excesiva dependencia y confianza en la AIG es una forma de riesgo o amenaza que puede describirse con base en dos parámetros: dependencia y confianza, y que se produce cuando estos superan un umbral en la interacción usual de los usuarios con la tecnología, de la misma forma en que una afición se transforma en una adicción. En el primer parámetro, enfocado en la dependencia, el riesgo surge cuando los y las estudiantes —pero también los profesores y las profesoras— adquieren conductas en el proceso de aprendizaje que transforman a las tecnologías de IAG en elementos imprescindibles del proceso, ante cuya ausencia el desarrollo de sus acciones propias no pueda llevarse a cabo. En el segundo parámetro, el de la confianza, el riesgo se manifiesta cuando el usuario adquiere la convicción de que la información o los contenidos generados por la IAG son necesariamente autosuficientes y correctos, renunciando a una posición crítica sobre ellos.

Aunque las tecnologías de IA son cautivadoras, tanto porque sorprenden como porque facilitan el trabajo, no debe olvidarse que los contenidos generados por la IAG no están —ni pueden estar— libres de errores u omisiones. Se ha detectado que ChatGPT presenta fallas de sentido

249

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIALRISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

común (Ji *et al.*, 2023) y no está aún preparado para proporcionar pruebas matemáticas o cálculos de alta calidad de manera consistente (Frieder *et al.*, 2023). Los errores en la IAG corresponden a los fenómenos que los expertos denominan «alucinación» —información generada sin base fáctica— y «divergencia» —fragmentos desalineados de la información de entrenamiento— (Xiao & Wang, 2021; Rebuffel *et al.*, 2022, p. 319; Bang *et al.*, 2023; Ji *et al.*, 2023). Los errores y los déficits de confiabilidad que presenta la IAG, como en el caso de ChatGPT, no la transforman en una herramienta inútil, a juicio de Dale (2021), sino que todo depende de para qué se use.

En consecuencia, así como no puede pretenderse, por ejemplo, obtener del *chatbot* consejos médicos porque sus errores pueden poner en riesgo la salud de personas (Dale, 2021), apoyar todo el proceso de enseñanza en esas herramientas también es desaconsejable. La excesiva dependencia y confianza en la IAG se presenta como un riesgo para los valores de honestidad y confianza que configuran la integridad académica. Si los estudiantes desarrollan una relación de dependencia con la IAG al nivel descrito, están renunciando al proceso mismo de aprendizaje y ello puede conducir a actos que afecten la honestidad, no solo por la posible realización de las conductas disvaloradas como la simulación de autoría, sino también por la misma renuncia de los estudiantes a su rol de protagonistas de su propio aprendizaje. En consecuencia, el resto de la comunidad académica, y principalmente los profesores, dejarán de confiar en sus estudiantes sobre el grado de aprendizaje que estos estén adquiriendo, lo que afectaría a la confianza como valor. Indirectamente, ello puede conducir a afectaciones a valores de la integridad académica como la justicia y el respeto.

V.1.2. Irrealizabilidad del proyecto pedagógico

La utilización masiva, descontrolada y sin supervisión de IAG puede poner en riesgo la «realizabilidad del proyecto pedagógico» de una institución de educación superior. Este riesgo es especialmente fuerte en aquellas instituciones que han optado por modelos pedagógicos basados en el desarrollo de competencias o de resultados de aprendizaje. El mal uso (y abuso) de las tecnologías de IAG puede llevar a que los estudiantes no logren el desarrollo de las competencias o de los aprendizajes esperados. En un escenario —espero— distópico en esta materia, los estudiantes pueden transformarse en meros operadores de máquinas generadoras de información, pero sin que logren una comprensión ni un análisis crítico, incapaces de construir su propio conocimiento y de generar otros nuevos.

Este riesgo de irrealizabilidad amenaza indirectamente los valores de la integridad académica, ya que una institución incapaz de concretar sus propios objetivos en docencia podría generar una cultura en la que nada

importe, con los valores de la integridad académica como los primeros candidatos a dicha indiferencia.

V.I.3. Incorporación de sesgos y discriminación en el aprendizaje

Uno de los principales problemas de la IA en general, y en particular de las tecnologías generativas, es la existencia en ellas de sesgos y criterios de discriminación —arbitraria— provenientes de la información utilizada en sus procesos de entrenamiento (Bender *et al.*, 2021, pp. 613 y ss.; Weidinger *et al.*, 2021, pp. 9-18) y de su propio desarrollo. Así, la inclusión y la diversidad son factores que se consideran necesarios para el desarrollo de una IA ética (Dignum, 2019, pp. 100-101). Los sesgos y la discriminación son los presentes en la sociedad y en la información con la que son entrenadas las herramientas de IA, o los de los propios programadores de las aplicaciones (Whitby, 2011, p. 143). El uso de tales herramientas, en consecuencia, produce el riesgo de que los estudiantes incorporen en su proceso de aprendizaje tales sesgos y criterios de discriminación. Es cierto que las visiones sesgadas y discriminatorias parecen ser propias del ser humano, lo que significa que tales sesgos y discriminaciones también estarán presentes en la enseñanza y el aprendizaje sin el uso de IAG. Pero si esa condición es propiamente humana, parece que estamos en la necesidad de tolerar ciertos márgenes «naturales» de sesgos y discriminación, al punto de que hemos desarrollado mecanismos para detectarlos y, en lo posible, superarlos. Una cuestión distinta es que sea conveniente que aceptemos —o tengamos que aceptar— los que provienen de máquinas. El riesgo se hace más intenso si se considera el *black box effect* propio de la IA, una característica de complejidad de las tecnologías de *deep learning* que hace imposible reconstruir los procesos que ejecutó la máquina para llegar al resultado que entrega, lo que prácticamente atribuye a sus resultados un carácter inescrutable. «El peligro está en crear y usar decisiones que no son justificables, legitimables o que simplemente no permiten obtener explicaciones detalladas de ese comportamiento» ejecutado por la máquina (Barredo Arrieta *et al.*, 2020, p. 83).

El aprendizaje mediado por errores, sesgos o discriminaciones afecta directamente la honestidad. Los miembros de la comunidad académica que aprenden bajo esas condicionantes no podrán ser honestos y, de hecho, ni siquiera estarán en condiciones de percibir sus propios errores. Pero también amenaza la confianza entre los distintos miembros de la comunidad, presentando un escenario propicio para una pérdida del respeto, otro de los valores amenazados.

251

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIAL

RISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

V.2. Riesgos de resultado individuales

V.2.1. Responsabilidad legal

Dentro de los riesgos de resultado de carácter individual se encuentra el de sufrir consecuencias o responsabilidades legales.

Las conductas disvaloradas antes mencionadas (plagio y suplantación y simulación de autoría) pueden llevar a que la persona que incurra en ellas infrinja normas jurídicas universitarias o administrativas por las que puede ser sancionada, produzca un daño a terceros que deba indemnizar o cometa un delito que le haga penalmente responsable. Asimismo, esas conductas pueden llegar a comprometer la responsabilidad civil o administrativa de la propia institución universitaria por infringir los deberes de supervisión sobre sus propios estudiantes al permitir que esas conductas disvaloradas se ejecuten o por no adoptar medidas de prevención adecuadas para impedir las.

La presencia de conductas disvaloradas pueden hacer que sus protagonistas intenten evadir los mecanismos de control para evitar sufrir las consecuencias legales. El éxito de estas maniobras de evasión afecta directamente los valores académicos de justicia y responsabilidad. Por otro lado, un sistema eficaz de control y supervisión que sea capaz de reprimir estas conductas puede conducir a otros riesgos, como la irrealizabilidad del modelo pedagógico.

V.2.2. Aprendizaje distorsionado y adquisición de sesgos y discriminaciones

De igual forma, el mal uso de herramientas de IA puede conducir a un aprendizaje distorsionado, ya sea porque el o la estudiante adquirió información incorrecta, sesgada o discriminatoria; porque no internalizó esa información y, por tanto, es incapaz de procesarla; o porque las mejoras que el o la estudiante introdujo pueden no reflejar su conocimiento como autor o autora de un texto escrito y corregido automáticamente (Perkins, 2023). La sobreutilización de herramientas de IA como fuente de información o como instrumentos para «hacer el trabajo», la no contrastación con otras fuentes y la ausencia de análisis crítico pueden llevar a que el aprendizaje logrado se encuentre completamente distorsionado por los sesgos y la discriminación presente en la IA.

Cabe subrayar que un aprendizaje distorsionado o sesgado afecta directamente los valores de integridad académica de honestidad, confianza y respeto.

V.2.3. Sobrecalificación infundada

El uso de tecnologías algorítmicas de generación de texto y contenido puede conducir a que los ensayos, memorias o tesis —en general, las evaluaciones escritas— arrojen un rendimiento asimétrico con las

competencias desarrolladas o los aprendizajes obtenidos; en otras palabras, a una sobrecalificación que carezca de fundamento fáctico en términos del real aprendizaje del o la estudiante.

La conducta de mejorar el rendimiento numérico en las evaluaciones —notas o calificaciones— usando deshonestamente la tecnología introduce, como es evidente, un elemento de injusticia. Tales conductas aumentan la probabilidad de una mejor calificación con una fracción del esfuerzo que otro/a estudiante honesto/a puso en la realización de la misma tarea sometida a evaluación. El subsidio que el uso deshonesto de la tecnología puede producir es una amenaza al valor de justicia que forma parte de la definición de integridad académica. Aunque el proceso de evaluación es el mismo para todos los estudiantes, el uso fraudulento de tecnología no está considerado dentro de los criterios evaluativos y hace que se obtenga la misma evaluación —o mejor— con menos esfuerzo que quien se empeñó más.

V.3. Riesgos de resultado colectivos

V.3.1. Pérdida de competitividad de la institución

En la esfera de los riesgos de resultado de carácter colectivo, la utilización indebida de tecnologías de IA puede llegar a amenazar la competitividad de una institución de educación de superior. Una universidad en la que el uso de tales prácticas incorrectas se masifique y no se adopten medidas de mitigación puede ver afectada la competitividad de sus titulaciones, tanto por una cuestión de imagen pública como por la menor calidad de los profesionales que egresen de ella.

V.3.2. Contaminación de la cultura académica

La difusión de prácticas deshonestas y su falta de represión puede hacer que estas terminen naturalizándose y normalizándose. Este proceso sostenido en el tiempo, que permite que las prácticas deshonestas se naturalicen y normalicen, deteriora y contamina la cultura académica de una institución universitaria. La contaminación se manifiesta en el desapego de sus miembros a los valores de integridad académica y, con ello, en la renuncia a los valores que esta implica.

V.3.3. Degradación del valor ético de la institución

El proceso de degradación a que puede conducir una cultura académica contaminada puede generar, finalmente, una pérdida del valor ético de tal institución. El valor ético —esto es, el fuerte compromiso real y material con la integridad académica— es una condición que debería ser considerada como el principal activo intangible en una universidad. Así, la pérdida de competitividad y la degradación de su valor ético muestra un escenario profundamente perverso, caracterizado por la negación de la integridad académica.

253

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIALRISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

VI. MEDIDAS DE MITIGACIÓN

Los desafíos que plantea la IA²⁹ en la actividad docente universitaria conducen la mirada a lo que se ha descrito como la destreza digital (*digital dexterity*); esto es, la «capacidad de autoorganizarse rápidamente para ofrecer un nuevo valor a partir de las tecnologías digitales» (Soule *et al.*, 2016, p. 5). Me parece razonable que, frente a las nuevas realidades que se están construyendo a partir de los desarrollos tecnológicos, las universidades adopten una actitud de reconocimiento hacia ellas e inicien un proceso de reflexión que les permita obtener los beneficios que tales tecnologías ofrecen, aprovechándolos para diseñar y aplicar en forma efectiva medidas que mitiguen los riesgos que ellas plantean. Estas tareas deberían ser asumidas con un sentido de urgencia, atendida la velocidad con la que la tecnología se desarrolla³⁰.

En este sentido, creo necesario explorar cuatro medidas de mitigación de los riesgos antes descritos: la adaptación de marcos normativos, la revalorización del pensamiento crítico, los cambios en los paradigmas de la docencia y el cuestionamiento de los modelos de *e-learning*. A continuación, se ofrece una breve explicación de ellas.

VI.1. Adaptación de los marcos normativos

El normativo es uno de los elementos que conforman una organización. Las normas son mandatos de comportamiento que poseen ciertas características que las hacen importantes, a saber: a) tienen la capacidad de aumentar la predictibilidad de los comportamientos de los otros —que no es lo mismo que pretender que las normas *determinen* comportamientos ajenos—; b) guían la selección de comportamientos que efectúan las personas, ya que las normas indican qué es lo esperable y qué no; c) obligan a mejorar la coordinación entre personas colaboradoras en un proyecto o tarea común (Malle *et al.*, 2017, pp. 4-5); y d) tienen una cierta capacidad para motivar el comportamiento de las personas, mostrando la trascendencia de los intereses relevantes a proteger, en los que se denominan los «efectos expresivo-integradores» de las normas (Díez Ripollés, 2003, p. 151).

La adaptación de los marcos normativos como medida de mitigación consiste en que las universidades inicien un proceso interno de discusión que lleve a una adaptación de sus normas internas a las nuevas

29 A modo de ejemplo, Sparrow (2022) y Cassidi (2023) destacan los problemas que las tecnologías generativas basadas en IA suponen para los sistemas de evaluación en las universidades, mostrando cómo casas de estudio universitarias británicas y australianas han adoptado medidas para impedir las conductas deshonestas en tales evaluaciones.

30 Stephen Marche (2022, citado por Rudolph *et al.*, 2023) ha estimado un horizonte temporal de diez años para que la academia enfrente estas nuevas realidades construidas a partir de avances tecnológicos: «dos años para que los estudiantes descubran la tecnología, tres años más para que los profesores reconozcan que los estudiantes están usando la tecnología y luego cinco años para que los administradores de la universidad decidan qué hacer, si acaso, al respecto (p. 13)».

realidades marcadas por la IA. Se trata de que las reglamentaciones estén construidas considerando estos nuevos riesgos para evitar vacíos normativos que hagan que, frente a una conducta disvalorada, esta deba ser sancionada aun en ausencia de norma —afectando la predictibilidad requerida de las normas— o que no pueda ser sancionada por ausencia de normativa que lo permita, afectando la vigencia de valores de integridad académica. En este sentido, la adaptación de los marcos normativos pasa por la incorporación de una regulación especialmente diseñada para lo que aquí he denominado «simulación de autoría».

Además de verificar que las normas contemplen los nuevos escenarios propuestos por las tecnologías, el proceso de definir estas nuevas reglas puede ser también provechoso para la comunidad académica. En efecto, el proceso de construir un nuevo cuerpo regulatorio requiere de, al menos, tres elementos: que las personas conozcan el fenómeno que se pretende regular, tanto en su descripción como en sus aspectos valorativos; que esté claro cuáles son las opciones para regular ese fenómeno; y un consenso sobre la respuesta que las normas van a proporcionar frente a ese fenómeno y las consecuencias que se van a generar ante su incumplimiento. La concreción de esos tres elementos en la realidad es un proceso de por sí provechoso porque estos implican el compromiso de los miembros de la comunidad académica con el problema y su posible o posibles soluciones.

VI.2. Revalorización del pensamiento crítico

La recopilación de información, el levantamiento bibliográfico e, incluso, la síntesis de conocimientos existentes son tareas que pueden ser ejecutadas automatizadamente con la ayuda de tecnologías de IA y, en particular, las generativas. Esto significa que tales tareas deberían dejar de tener una incidencia alta en el proceso de evaluación. Pero hay un área en la que, por el momento, las aplicaciones de IAG no han sido capaces de igualarnos: el pensamiento crítico y reflexivo. Si una parte importante de las tareas que integran la actividad de los estudiantes puede ser externalizada, lo lógico sería concentrarse en aquellas que aún no pueden serlo. Esta es una de las externalidades positivas que plantean las tecnologías de IAG: ofrecen una buena razón para revalorizar el pensamiento crítico.

Pero, para revalorizar el pensamiento crítico, primero se debe tener un concepto sobre él. De acuerdo con la descripción de Paul (2005), el pensamiento crítico es aquel que comprende el análisis, la evaluación y la mejora del propio pensamiento, y su resultado es el pensamiento creativo (p. 28).

Orientar la docencia hacia el pensamiento crítico requiere, después de la definición, de un consenso sobre la forma en la que dicha orientación

255

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIALRISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

puede ser materializada y de las metodologías para lograrlo. Todas estas cuestiones deberían estar en el centro de la actividad académica en las universidades.

La adopción del pensamiento crítico como un estándar para la evaluación, por ejemplo, ayuda a mitigar los riesgos de la IAG. Si la docencia y la evaluación se quedan en los niveles de datos, repetición y memorización, tales actividades pueden sucumbir frente a las tecnologías de la IA.

VI.3. Cambios en los paradigmas de la docencia y la evaluación

La revalorización del pensamiento crítico como horizonte de la docencia produce como consecuencia que una revisión, también crítica, de los paradigmas de docencia y evaluación se torne necesaria. Esta medida de mitigación consiste en la generación de espacios de discusión académica en la comunidad universitaria sobre los beneficios y riesgos de la IA en la actividad docente. Este proceso de discusión tiene como externalidad la difusión de las problemáticas y pone a los miembros de la comunidad académica en la posición de tener que analizarlos y adoptar puntos de vista y convicciones sobre la cuestión. El proceso de discusión puede generar una simbiosis importante —y positiva— con la medida de adaptación normativa, ya que el mayor conocimiento de la norma asegura su mayor efecto expresivo-integrador.

A nivel operativo, la irrupción de la IAG obligará a repensar aspectos importantes de la docencia, como la evaluación³¹. La IAG puede ayudar enormemente en el proceso de evaluación, proveyendo a los profesores de instrumentos creativos capaces de estimular el interés y la imaginación de los estudiantes, pero sus riesgos obligan a adoptar medidas más concretas. Por lo pronto, un desafío para los profesores es el de ser capaces de generar «preguntas resistentes al fraude» (Susnjak, 2022, p. 4), como aquellas en las que el estudiante deba generar un conocimiento nuevo aplicando los contenidos estudiados. Mi área de especialización, el derecho penal, se presta particularmente para preguntas de desarrollo de casos, por ejemplo. Otras cuestiones, como la revalorización de las exámenes presenciales, aparecen como buenas alternativas, así como diseñar instrumentos de evaluación que escapen de las posibilidades —actuales— de la tecnología: preguntas con enunciados largos cuya resolución requiera aplicación o pensamiento crítico, de la habilidad de integrar conocimientos o que se refieran a información que quede fuera del periodo comprendido en la información de entrenamiento (Rudolph *et al.*, 2023). En otras

³¹ Al respecto, pueden consultarse los interesantes trabajos contenidos en la obra colectiva editada por Margaret Bearman *et al.* (2020).

palabras, frente a la llegada de las tecnologías de IA, la evaluación —más que nunca antes— requiere apuntar a la recogida de evidencia sobre si los y las estudiantes fueron capaces de internalizar los conceptos marco del derecho, como el razonamiento jurídico o el Estado de derecho, y su integración con los conceptos centrales de cada disciplina, de acuerdo con la distinción entre ambos propuesta por Donson y O'Sullivan (2017, pp. 23-26).

VI.4. Cuestionamiento de modelos de *e-learning*

Los entornos de aprendizaje asincrónico, también conocidos como *e-learning*, ofrecen buenas oportunidades para una difusión de prácticas desleales y conductas disvaloradas con relación a la integridad académica, en especial la evaluación en tales entornos asincrónicos (Grijalva *et al.*, 2006; Stuber-McEwen *et al.*, 2009; Carabantes Alarcón, 2020; Holden *et al.*, 2021; Chiang *et al.*, 2022). La creciente popularidad de estos entornos de aprendizaje debería considerar la irrupción de las tecnologías de IAG. Una valoración profunda de la compatibilidad de ambas es, en todo caso, recomendable, en particular en los procesos de evaluación. En estos entornos, la evaluación carece de los mecanismos de control formal que conllevan las aplicadas presencialmente. Uno de esos mecanismos —el más básico— es la vigilancia directa, que impide que los estudiantes externalicen el desarrollo de sus respuestas a aplicaciones de AIG. Si el entorno carece de esa vigilancia directa, el desarrollo de una evaluación puede ser externalizada a una IAG. La complejidad del tema impide analizarlo en profundidad aquí, por lo que me parece apropiado solo dejarlo planteado para futuras investigaciones sobre la materia.

VII. CONCLUSIONES

El vertiginoso desarrollo tecnológico que está registrando la IA tiene un enorme potencial para la docencia en las universidades. En el otro lado de la moneda, esas mismas tecnologías plantean desafíos sobre los que se requieren procesos continuos de investigación, discusión y análisis. En particular, las tecnologías algorítmicas de escritura y de generación de contenido (IAG) están obligando —queramos reconocerlo o no— a replantear los paradigmas con los que hemos impartido la docencia y, como parte de ella, aquellos con los que se ha evaluado el aprendizaje o la obtención de competencias por parte de los estudiantes.

Una buena demostración de las nuevas realidades que plantea la IA son las posibilidades que ofrece la aplicación ChatGPT. Usada deshonestamente, esta puede suplantar al estudiante en sus evaluaciones, principalmente en las escritas por medio de ensayos, memorias de grado o tesis.

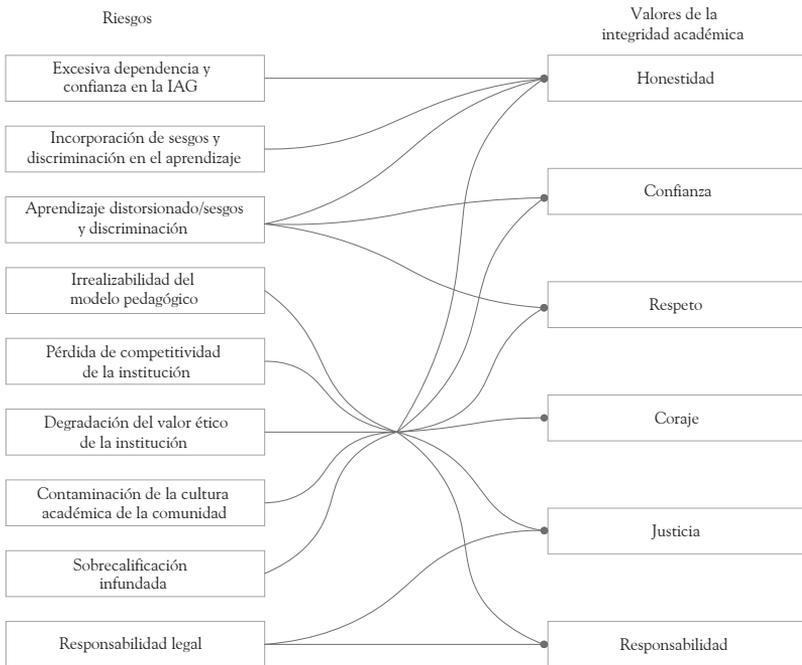
257

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIALRISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

Este artículo ha tenido como objetivo la exposición sintética de las relaciones eventualmente problemáticas con la IA (y, dentro de ella, con la IAG). En ese sentido, actúa como un primer encuentro con la temática expuesta y, a su vez, representa el inicio de una línea de investigación sobre la materia, intentando un examen de la bibliografía existente. Su hipótesis, que ha tomado como punto de partida la definición de integridad académica como un conjunto de valores (ver figura 1), plantea que las tecnologías de IAG amenazan directa o indirectamente tales valores y se configuran, en consecuencia, como riesgos para esta.

Como parte del análisis, se ha propuesto una taxonomía de riesgos, subclasificada en riesgos procedimentales y de resultado, y en su exposición se ha establecido la relación entre tales riesgos y los valores de la integridad académica. A modo de síntesis, dichas relaciones se grafican en la figura 3.

Figura 3. Relación entre los riesgos para la integridad académica y los valores que la componen



Fuente: elaboración propia.

Finalmente, y sobre la base de la identificación de los riesgos y de sus relaciones conflictivas con los valores de la integridad académica, se propusieron cuatro medidas de mitigación. Estas son la adaptación de los marcos normativos de las universidades para que sus propias

regulaciones den cuenta de las tecnologías de IAG y de su impacto sobre la actividad académica; la revalorización del pensamiento crítico como estrategia para relevar la importancia de la docencia y cómo esta debería enfocarse al desarrollo en los y las estudiantes de una actitud crítica y reflexiva, habilidades que, por el momento escapan a las posibilidades de la IAG; cambios en los paradigmas de docencia y evaluación, lo que supone reconocer la existencia de la IAG y cómo esta pueda alterar negativamente tales procesos; y el cuestionamiento de los modelos pedagógicos asincrónicos basados en tecnologías de información y comunicación.

La llegada de las tecnologías de IA y, en particular, de las de IAG, requiere que las universidades partan por conocerlas y comprenderlas. Este proceso permitirá tomar conciencia no solo de las ventajas y posibilidades que ellas entregan para la innovación pedagógica, sino también de los riesgos. La docencia, entendida como un proceso, podría verse profundamente alterada por la irrupción tecnológica. Como las universidades no tienen capacidad ni de detener el desarrollo tecnológico ni de impedir que sus estudiantes utilicen estas tecnologías emergentes, solo les queda actuar para aprovecharlas y regularlas. En todo caso, ellas ofrecen un buen escenario para iniciar o retomar un proceso de discusión sobre la integridad académica.

REFERENCIAS

Abbott, R., & Sarch, A. (2019). Punishing Artificial Intelligence: Legal Fiction or Science Fiction. *UC Davis Law Review*, 53(1), 323-384.

Abbott, R., & Sarch, A. (2020). Punishing artificial intelligence: Legal fiction or science fiction. En S. Deakin y C. Markou (eds.), *Is Law Computable? Critical perspectives on Law and Artificial Intelligence* (pp. 177-204). Hart Publishing. <https://doi.org/doi.org/10.5040/9781509937097.ch-008>

Abdelaal, E., Walpita Gamage, S., & Mills, J. (2019). Artificial Intelligence Is a Tool for Cheating Academic Integrity. *AAEE 2019 Annual Conference*.

Abukmeil, M., Ferrari, S., Genovese, A., Piuri, V., & Scotti, F. (2021). A Survey of Unsupervised Generative Models for Exploratory Data Analysis and Representation Learning. *ACM Computing Surveys*, 54(5), 1-40. <https://doi.org/https://doi.org/10.1145/3450963>

Aitchison, C., & Mowbray, S. (2016). Doctoral Writing Markets: Exploring the Grey Zone. En T. Bretag (ed.), *Handbook of Academic Integrity* (pp. 287-301). Springer.

Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, 97(22), 3242-3250.

259

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIALRISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

Alibašić, H., & Rose, J. (2019). Fake News in Context: Truth and Untruths. *Public Integrity*, 21(5), 463-468. <https://doi.org/https://doi.org/10.1080/10999922.2019.1622359>

Aljanabi, M., & ChatGPT. (2023). ChatGPT: Future Directions and Open possibilities [editorial]. *Mesopotamian Journal of Cybersecurity*, 2023, 16-17. <https://doi.org/https://doi.org/10.58496/MJCS/2023/003>

Aydin, Ö., & Karaarslan, E. (2022). OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare. https://papers.ssm.com/sol3/papers.cfm?abstract_id=4308687. <https://doi.org/https://dx.doi.org/10.2139/ssrn.4308687>

Azaria, A. (2022). ChatGPT Usage and Limitations. *HAL Open Science*, 1-6. <https://doi.org/https://doi.org/10.31219/osf.io/5ue7n>

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *ArXiv, abs/2302.0*.

Barfield, W., & Pagallo, U. (2020). *Advanced Introduction to Law and Artificial Intelligence*. Edward Elgar Publishing. <https://doi.org/https://doi.org/10.4337/9781789905137>

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.

Bartz, D. (2023). As ChatGPT's popularity explodes, U.S. lawmakers take an interest. *Reuters*. <https://www.reuters.com/technology/chatgpts-popularity-explodes-us-lawmakers-take-an-interest-2023-02-13/>

Bearman, M., Dawson, P., Ajjawi, R., Tai, J., & Boud, D. (eds.). (2020). *Re-imagining University Assessment in a Digital World*. Springer.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? En *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). <https://doi.org/https://doi.org/10.1145/3442188.3445922>

Boddington, P. (2017). *Towards a Code of Ethics for Artificial Intelligence*. Springer. <https://doi.org/https://doi.org/10.1007/978-3-319-60648-4>

Bommarito, M. J., & Katz, D. M. (2022). GPT Takes the Bar Exam. *ArXiv, abs/2212.1*. <https://doi.org/https://doi.org/10.2139/ssrn.4314839>

Bretag, T. (ed.) (2016). *Handbook of Academic Integrity*. Springer. <https://doi.org/https://doi.org/10.1007/978-981-287-098-8>

Brown, W., & Fleming, D. H. (2020). Celebrity headjobs: or oozing squid sex with a framed-up leaky {Schar-JØ}. *Pom Studies*, 7(4), 357-366. <https://doi.org/10.1080/23268743.2020.1815570>

Bryson, J. J. (2020). The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation. En M. Dubber, F. Pasquale y S. Das (eds.), *The Oxford Handbook of Ethics of AI* (pp. 3-25). Oxford University Press. <https://doi.org/https://doi.org/10.1093/oxfordhb/9780190067397.001.0001>

Caldera, E. (2019). "Reject the Evidence of Your Eyes and Ears": Deepfakes and the Law of Virtual Replicants. *Seton Hall Law Review*, 50(1), 177-205.

Carabantes Alarcón, D. (2020). Integridad académica y educación superior: nuevos retos en la docencia a distancia. *Análisis Carolina*, 28, 1-13. https://doi.org/https://doi.org/10.33960/ac_38.2020

Carrasco, J. P., García, E., Sánchez, D. A., Estrella Porter, P. D., De La Puente, L., Navarro, J., & Cerame, A. (2023). ¿Es capaz "ChatGPT" de aprobar el examen MIR de 2022? Implicaciones de la inteligencia artificial en la educación médica en España. *Revista Española de Educación Médica*, 1, 55-69. <https://doi.org/https://doi.org/10.6018/edumed.556511>

Cassidy, C. (2023, 10 de enero). Australian universities to return to 'pen and paper' exams after students caught using AI to write essays. *The Guardian*. <https://www.theguardian.com/australia-news/2023/jan/10/universities-to-return-to-pen-and-paper-exams-after-students-caught-using-ai-to-write-essays>

Chace, C. (2018). *Artificial intelligence and the two singularities*. CRC Press. <https://doi.org/https://doi.org/10.1201/9781351254465>

Chesney, R., & Citron, D. K. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107, 1753-1819. <https://doi.org/https://doi.org/10.2139/ssrn.3213954>

Chiang, F.-K., Zhu, D., & Yu, W. (2022). A systematic review of academic dishonesty in online learning environments. *Journal of Computer Assisted Learning*, 38(4), 907-928. <https://doi.org/https://doi.org/10.1111/jcal.12656>

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. (2023). ChatGPT Goes to Law School [white paper]. *Journal of Legal Education*, (387). <https://doi.org/https://dx.doi.org/10.2139/ssrn.4335905>

Chomanski, B. (2022). Legitimacy and automated decisions: the moral limits of algocracy. *Ethics and Information Technology*, 24(3), 1-9. <https://doi.org/https://doi.org/10.1007/s10676-022-09647-w>

Comas Forgas, R., & Sureda Negre, J. (2008). El intercambio y compra-venta de trabajos académicos a través de Internet. *Revista Electrónica de Tecnología Educativa*, (26), 1-16. <https://doi.org/https://doi.org/10.21556/edutec.2008.26.466>

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 0(0), 1-12. <https://doi.org/https://doi.org/10.1080/14703297.2023.2190148>

Council of Europe. (2018). *Study on the Human Rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory*

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIAL

RISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

implications. Council of Europe. <https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5>

Crawford, J., Cowling, M., & Allen, K.-A. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching & Learning Practice*, 20(1), 1-19. <https://doi.org/https://doi.org/10.53761/1.20.3.02>

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering*, 27(1), 113-118. <https://doi.org/https://doi.org/10.1017/s1351324920000601>

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, 29(3), 245-268. <https://doi.org/https://doi.org/10.1007/s13347-015-0211-11>

Danaher, J. (2022). Freedom in an Age of Algocracy. En S. Vallor (ed.), *The Oxford Handbook of Philosophy of Technology* (pp. 250-272). Oxford University Press. <https://doi.org/https://doi.org/10.1093/oxfordhb/9780190851187.013.16>

Dawson, P. (2021). *Defending Assessment Security in a Digital World Preventing E-Cheating and Supporting Academic Integrity in Higher Education*. Routledge. <https://doi.org/https://doi.org/10.4324/9780429324178>

Departamento de Evaluación, Medición y Registro Educacional (Demre). (2022). *Modelo de Prueba de Historia y Ciencias Sociales*. <https://demre.cl/publicaciones/2023/2023-22-03-31-modelo-historia-csociales>

Díez Ripollés, J. L. (2003). El Derecho penal simbólico y los efectos de la pena. En L. Arroyo Zapatero, U. Neumann y A. Nieto Marín (eds.), *Crítica y justificación del Derecho penal en el cambio de siglo* (pp. 147-172). Ediciones de la Universidad de Castilla-La Mancha.

Dignum, V. (2019). *Responsible artificial intelligence. How to develop and use AI in a responsible way*. Springer. <https://doi.org/https://doi.org/10.1007/978-3-030-30371-6>

Donath, J. (2020). Ethical Issues in Our Relationship with Artificial Entities. En M. D. Dubber, F. Pasquale y S. Das (eds.), *The Oxford Handbook of Ethics of AI* (pp. 53-73). Oxford University Press. <https://doi.org/https://doi.org/10.1093/oxfordhb/9780190067397.013.3>

Donsion, F., & O'Sullivan, C. (2017). Building blocks or stumbling block? Teaching actus reus and mens rea in criminal law. En K. Gledhill y B. Living (eds.), *The Teaching of Criminal Law. The pedagogical imperatives* (pp. 21-33). Routledge.

Dowling, M., & Lucey, B. (2023). ChatGPT for (Finance) research: The Bananarama Conjecture. *Finance Research Letters*, 53. <https://doi.org/https://doi.org/10.1016/j.frl.2023.103662>

Echavarría, M. A. (2014). ¿Qué es el plagio? Propuesta conceptual del plagio punible. *Revista Facultad de Derecho y Ciencias Políticas*, 44(121), 699-720.

Echavarría, M. A. (2016). El delito de plagio: una propuesta de regulación penal de la infracción al derecho de autor. *Cuadernos de Derecho Penal*, 15, 85-101. <https://doi.org/https://doi.org/10.22518/20271743.577>

Eitel-Porter, R. (2021). Beyond the promise: implementing ethical AI. *AI and Ethics*, 1(1), 73-80. <https://doi.org/https://doi.org/10.1007/s43681-020-00011-6>

Eke, D. O. (2023). ChatGPT and the rise of generative AI: Threat to academic integrity? *Journal of Responsible Technology*, 13. <https://doi.org/https://doi.org/10.1016/j.jrt.2023.100060>

Faraldo-Cabana, P. (2015). Who Dares Fine a Murderer? The Changing Meaning of Money and Fines in Western European Criminal Systems. *Social & Legal Studies*, 25(4), 489-507. <https://doi.org/https://doi.org/10.1177/0964663915618545>

Faraldo-Cabana, P. (2018). Research excellence and Anglophone dominance: The case of law, criminology and social science. En K. Carrington, R. Hogg, J. Scott y M. Sozzo (eds.), *The Palgrave Handbook of Criminology and the Global South* (pp. 163-182). Palgrave. https://doi.org/https://doi.org/10.1007/978-3-319-65021-0_9

Faraldo-Cabana, P. (2019). Consecuencias imprevistas de la dominación anglófona en las ciencias sociales y jurídicas. *RES. Revista Española de Sociología*, 28(1), 45-60. <https://doi.org/https://doi.org/10.22325/fes/res.2018.57>

Faraldo-Cabana, P., & Lamela, C. (2021). How international are the top international journals of criminology and criminal justice? *European Journal on Criminal Policy and Research*, 27(2), 151-174. <https://doi.org/https://doi.org/10.1007/s10610-019-09426-2>

Felländer, A., Rebane, J., Larsson, S., Wiggberg, M., & Heintz, F. (2022). Achieving a Data-Driven Risk Assessment Methodology for Ethical AI. *Digital Society*, 1(2), 13. <https://doi.org/https://doi.org/10.1007/s44206-022-00016-0>

Fierens, M., Rossello, S., & Wauters, E. (2021). Setting the scene: On AI ethics and regulation. En J. De Bruyne y C. Vanleenhove (eds.), *Artificial Intelligence and the Law* (pp. 49-72). Intersentia.

Fishman, T. (2016). Academic Integrity as an Educational Concept, Concern, and Movement in US Institutions of Higher Learning BT - Handbook of Academic Integrity. En T. Bretag (ed.), *Handbook of Academic Integrity* (pp. 7-21). Springer Singapore. https://doi.org/https://doi.org/10.1007/978-981-287-098-8_1

Floridi, L. (2023). AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models. *Philosophy & Technology*, 36(1), 15. <https://doi.org/https://doi.org/10.2139/ssrn.4358789>

Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P.C., Chevalier, A., & Berner, J. J. (2023). Mathematical Capabilities of ChatGPT. *ArXiv, abs/2301.1*.

García-Peñalvo, F. (2023). La percepción de la Inteligencia Artificial en contextos educativos tras el lanzamiento de ChatGPT: disrupción o pánico. *Education in the Knowledge Society*, 24. <https://doi.org/https://doi.org/10.14201/eks.31279>

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIAL

RISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

- García-Ull, F. (2021). Deepfakes: The next challenge in fake news detection. *Análisi*, 64, 103-120. <https://doi.org/https://doi.org/10.5565/rev/analisi.3378>
- García-Villegas, M., Franco-Pérez, N., & Cortés-Arbeláez, A. (2016). Perspectives on Academic Integrity in Colombia and Latin America. En T. Bretag (ed.), *Handbook of Academic Integrity* (pp. 161-180). Springer. https://doi.org/https://doi.org/10.1007/978-981-287-098-8_10
- Gordijn, B., & Have, H. ten. (2023). ChatGPT: evolution or revolution? *Medicine, Health Care and Philosophy*. <https://doi.org/https://doi.org/10.1007/s11019-023-10136-0>
- Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). ChatGPT is not all you need. A State of the Art Review of large Generative AI models. *ArXiv, abs/2301.0*.
- Grijalva, T. C., Nowell, C., & Kerkvliet, J. (2006). Academic Honesty and Online Courses. *College Student Journal*, 40(1), 180-185.
- Gritsenko, D., & Wood, M. (2020). Algorithmic governance: A modes of governance approach. *Regulation and Governance*. <https://doi.org/https://doi.org/10.1111/rego.12367>
- Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), 99-120. <https://doi.org/https://doi.org/10.1007/s11023-020-09517-8>
- Hagendorff, T. (2022). Blind spots in AI ethics. *AI and Ethics*, 2(4), 851-867. <https://doi.org/https://doi.org/10.1007/s43681-021-00122-8>
- Harris, D. (2019). Deepfakes: False Pornography Is Here and the Law Cannot Protect You. *Duke Law & Technology Review*, 17, 99-127.
- Heaven, D. (2023). GPT-4 is bigger and better than ChatGPT—but OpenAI won't say why. *MIT Technology Review*. https://www.technologyreview.com/2023/03/14/1069823/gpt-4-is-bigger-and-better-chatgpt-openai/?utm_source=engagement_email&utm_medium=email&utm_campaign=wklysun&utm_content=03.19.23.nonsubs_eng&mc_cid=d42ec9a242&mc_eid=a0446d3f11
- Helberger, N., & Diakopoulos, N. (2023). ChatGPT and the AI Act. *Internet Policy Review*, 12(1). <https://doi.org/https://doi.org/10.14763/2023.1.1682>
- Henry, J. V., & Oliver, M. (2022). Who Will Watch the Watchmen? The Ethico-political Arrangements of Algorithmic Proctoring for Academic Integrity. *Postdigital Science and Education*, 4(2), 330-353. <https://doi.org/https://doi.org/10.1007/s42438-021-00273-1>
- Hermansson, H., & Hansson, S. O. (2007). A three-party model for ethical risk analysis. *Risk Management*, 9(3), 129-144. <https://doi.org/https://doi.org/10.1057/palgrave.rm.8250028>
- Hern, A. (2023). TechScope: Will Meta's massive leak democratise AI – and at what cost? *The Guardian*. <https://www.theguardian.com/technology/2023/mar/07/techscope-meta-leak-llama-chatgpt-ai-crossroads>

Holden, O. L., Norris, M. E., & Kuhlmeier, V. A. (2021). Academic Integrity in Online Assessment: A Research Review. *Frontiers in Education*, 6. <https://doi.org/https://doi.org/10.3389/feduc.2021.639814>

Höppner, T., & Streatfeild, L. (2023). *ChatGPT, Bard & Co.: an introduction to AI for competition and regulatory lawyers*. Hausfeld. <https://www.hausfeld.com/en-us/what-we-think/competition-bulletin/chatgpt-bard-co-an-introduction-to-ai-for-competition-and-regulatory-lawyers/>

Hu, L. (2022). *Generative AI and Future*. GAN, GPT-3, DALL·E 2, and what's next. Medium. <https://pub.towardsai.net/generative-ai-and-future-c3b1695876f2>

International Center for Academic Integrity (ICAI). (2021). *The Fundamental Values of Academic Integrity* (3.^a ed.). https://academicintegrity.org/images/pdfs/20019_ICAI-Fundamental-Values_R12.pdf

Islam, I., & Islam, M. N. (2023). Opportunities and Challenges of ChatGPT in Academia: A Conceptual Analysis [preprints]. *Authorea*, 1-9. <https://doi.org/https://doi.org/10.22541/au.167712329.97543109/v1>

Jabotinsky, H. Y., & Sarel, R. (2023). *Co-authoring with an AI? Ethical dilemmas and Artificial Intelligence* [documento de trabajo]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4303959

Jelenic, M., & Kennette, L. N. (2022). Cheating: It depends how you define it. *Canadian Perspectives on Academic Integrity*, 5(2), 16-33. <https://doi.org/https://doi.org/10.11575/cpai.v5i2.75649> ISSN

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1-38. <https://doi.org/https://doi.org/10.1145/3571730>

Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2023). Is ChatGPT A Good Translator? A Preliminary Study. *ArXiv, abs/2301.0*. <https://doi.org/https://doi.org/10.48550/arXiv.2301.08745>

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Wellwe, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, 103. <https://doi.org/https://doi.org/10.35542/osf.io/5er8f>

Kirchner, J. H., Ahmad, L., Aaronson, S., & Leike, J. (2023). *New AI classifier for indicating AI-written text*. OpenAI.

Kleebayoon, A., & Wiwanitkit, V. (2023). Artificial Intelligence, Chatbots, Plagiarism and Basic Honesty: Comment. *Cellular and Molecular Bioengineering*. <https://doi.org/https://doi.org/10.1007/s12195-023-00759-x>

Lancaster, T. (2021). Academic Dishonesty or Academic Integrity? Using Natural Language Processing (NLP) Techniques to Investigate Positive Integrity in Academic Integrity Research. *Journal of Academic Ethics*, 19(3), 363-383. <https://doi.org/https://doi.org/10.1007/s10805-021-09422-4>

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIAL

RISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

Li, H., Moon, J. T., Purkayastha, S., Celi, L. A., Trivedi, H., & Gichoya, J. W. (2023). Ethics of large language models in medicine and medical research. *www.thelancet.com/digital-health*, 5, 333-335. [https://doi.org/https://doi.org/10.1016/S2589-7500\(23\)00083-3](https://doi.org/https://doi.org/10.1016/S2589-7500(23)00083-3)

Li, W., Wu, W., Chen, M., Liu, J., Xiao, X., & Wu, H. (2022). Faithfulness in Natural Language Generation: A Systematic Survey of Analysis, Evaluation and Optimization Methods. *ArXiv, abs/2203.0*.

Malle, B. F., Scheutz, M., & Austerweil, J. L. (2017). Networks of social and moral norms in human and robot agents. En M. I. Aldinhas Ferreira, E. E. Kadar, J. Silva Sequeira, G. S. Virk y M. O. Tokhi (eds.), *A world with robots. International Conference on Robot Ethics: ICRE 2015* (pp. 3-17). Springer. https://doi.org/https://doi.org/10.1007/978-3-319-46667-5_1

Melo, V. E. (2022). *Fake News*. Thomson Reuters, La Ley.

Microsoft. (2023). *Introducing Microsoft 365 Copilot—your copilot for work*. Microsoft. <https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/>

Milmo, D. (2023a). ChatGPT reaches 100 million users two months after launch. *The Guardian*. <https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app>

Milmo, D. (2023b). Google AI chatbot Bard sends shares plummeting after it gives wrong answer. *The Guardian*. <https://www.theguardian.com/technology/2023/feb/09/google-ai-chatbot-bard-error-sends-shares-plummeting-in-battle-with-microsoft>

Miquel-Vergés, J. (2022). Nuevas tecnologías en la educación superior virtual. Las tecnologías del ultrafalso y de la traducción cara a cara. *Human Review. Revista Internacional de Humanidades*, 12(4), 1-20. <https://doi.org/https://doi.org/10.37467/revhuman.v11.3967>

Moya, B. A., Eaton, S. E., Pethrick, H., Hayden, K. A., Brennan, R., Wiens, J., McDermott, B., & Lesage, J. (2023). Academic Integrity and Artificial Intelligence in Higher Education Contexts: A Rapid Scoping Review Protocol. *Canadian Perspectives on Academic Integrity*, 5(2), 59-75.

Newton, P. M. (2018). How Common Is Commercial Contract Cheating in Higher Education and Is It Increasing? A Systematic Review. *Frontiers in Education*, 3, 1-18. <https://doi.org/https://doi.org/10.3389/educ.2018.00067>

Newton, P.M., & Lang, C. (2016). Custom Essay Writers, Freelancers, and Other Paid Third Parties. En T. Bretag (ed.), *Handbook of Academic Integrity* (pp. 249-271). Springer. https://doi.org/https://doi.org/10.1007/978-981-287-098-8_38

OpenAI. (s.f.). *Chat GPT — Release Notes*. <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>

OpenAI. (2023). GPT-4 Technical Report. *ArXiv, abs/2303.0*. <https://api.semanticscholar.org/CorpusID:257532815>

Organización Mundial de la Propiedad Intelectual (OMPI). (1980). *Glosario de derecho de autor y derechos conexos*. Organización Mundial de Propiedad Intelectual.

Paul, R. (2005). The state of critical thinking today. *New Directions for Community Colleges*, (130), 27-38. <https://doi.org/https://doi.org/10.1002/cc.193>

Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice*, 20(2), 1-24. <https://doi.org/https://doi.org/10.53761/1.20.02.07>

Popova, M. (2020). Reading out of context: pornographic deepfakes, celebrity and intimacy. *Porn Studies*, 7(4), 367-381. <https://doi.org/10.1080/23268743.2019.1675090>

Powers, T. M., & Ganascia, J.-G. (2020). The Ethics of the Ethics of AI. En M. Dubber, F. Pasquale y S. Das (eds.), *The Oxford Handbook of Ethics of AI* (pp. 27-51). Oxford University Press. <https://doi.org/https://doi.org/10.1093/oxfordhb/9780190067397.013.2>

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is ChatGPT a General-Purpose Natural Language Processing Task Solver? *ArXiv*, *abs/2302.0*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Rahaman, S., Ahsan, T., Anjum, N., Rahman, M., & Rahman, N. (2023). The AI Race is On! Google's Bard and OpenAI's ChatGPT Head to Head: An Opinion Article. SSRN. <https://doi.org/http://dx.doi.org/10.2139/ssrn.4351785>

Raj, R. J. S., Babu, S. A., Jegatheesan, A., & Xavier, V. M. A. (2022). A GAN-Based Triplet FaceNet Detection Algorithm Using Deep Face Recognition for Autism Child. En J. D. Peter, A. H. Alavi y S. L. Fernandes (eds.), *Disruptive Technologies for Big Data and Cloud Applications. Proceedings of ICBDDC 2021* (pp. 177-187). Springer.

Rebuffel, C., Roberti, M., Soulier, L., Scoutheeten, G., Cancelliere, R., & Gallinari, P. (2022). Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*, 36(1), 318-354. <https://doi.org/10.1007/s10618-021-00801-4>

Riley, S., & Alvarez, L. C. (2023). *ChatGPT, Friend or Foe in the Classroom?* <https://otl.du.edu/chatgpt-friend-or-foe-in-the-classroom/>

Roe, J., & Perkins, M. (2022). What are Automated Paraphrasing Tools and how do we address them? A review of a growing threat to academic integrity. *International Journal for Educational Integrity*, 18(1), 15. <https://doi.org/https://doi.org/10.1007/s40979-022-00109-w>

Rogerson, A. M., & McCarthy, G. (2017). Using Internet based paraphrasing tools: Original work, patchwriting or facilitated plagiarism? *International Journal*

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIAL

RISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

for *Educational Integrity*, 13(1), 2. <https://doi.org/https://doi.org/10.1007/s40979-016-0013-y>

Rojas Chavarro, M., & Olarte Collazos, J. (2010). Plagio en el ámbito académico. *Revista Colombiana de Anestesiología*, 39(4), 537-538. <http://www.scielo.org.co/pdf/rca/v38n4/v38n4a10.pdf>

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching*, 6(1), 1-22. <https://doi.org/https://doi.org/10.37074/jalt.2023.6.1.9>

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4.ª ed.). Pearson.

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management (JDM)*, 31(2), 74-87. <https://doi.org/https://doi.org/10.4018/jdm.2020040105>

Simó Soler, E. (2023). Retos jurídicos derivados de la inteligencia artificial generativa. *Arxius*, 2. <https://doi.org/https://doi.org/10.31009/InDret.2023.i2.11>

Soule, D., Puram, A., Westerman, G., & Bonnet, D. (2016). *Becoming a Digital Organization: The Journey to Digital Dexterity* [documento de trabajo N.º 301]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2697688

Sparrow, J. (2022). 'Full-on robot writing': the artificial intelligence challenge facing universities. *The Guardian*. <https://www.theguardian.com/australia-news/2022/nov/19/full-on-robot-writing-the-artificial-intelligence-challenge-facing-universities>

Spivak, B., & Shepherd, S. (2021). Ethics, Artificial Intelligence, and Risk Assessment. *The Journal of the American Academy of Psychiatry and the Law*, 49. <https://doi.org/10.29158/JAAPL.210066-21>

Stuber-McEwen, D., Wiseley, P., & Hoggatt, S. (2009). Point, Click, and Cheat: Frequency and Type of Academic Dishonesty in the Virtual Classroom. *Online Journal of Distance Learning Administration*, 12(3).

Susnjak, T. (2022). ChatGPT: The End of Online Exam Integrity? *ArXiv*, [abs/2212.0](https://arxiv.org/abs/2212.0).

Talan, T., & Kalinkara, Y. (2023). The Role of Artificial Intelligence in Higher Education: ChatGPT Assessment for Anatomy Course. *International Journal of Management Information Systems and Computer Science*, 7(1), 33-40. <https://doi.org/http://dx.doi.org/10.33461/uybisbbd.1244777>

Taylor, L. (2023). Colombian judge says he used ChatGPT in ruling. *The Guardian*. <https://www.theguardian.com/technology/2023/feb/03/colombia-judge-chatgpt-ruling>

Tiku, N. (2022). The Google engineer who thinks the company's AI has come to life. *The Washington Post*. <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/>

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2022). An Introduction to Digital Face Manipulation. En C. Rathgeb, R. Vera-Rodriguez, R. Tolosana y C. Busch (eds.), *Handbook of Digital Face Manipulation and Detection. From DeepFakes to Morphing Attacks* (pp. 3-26). Springer.

Van der Sloot, B., & Wagenveld, Y. (2022). Deepfakes: regulatory challenges for the synthetic society. *Computer Law & Security Review*, 46. <https://doi.org/https://doi.org/10.1016/j.clsr.2022.105716>

Vanem, E. (2012). Ethics and fundamental principles of risk acceptance criteria. *Safety Science*, 50, 958-967. <https://doi.org/https://doi.org/10.1016/j.ssci.2011.12.030>

Weidinger, L., Mellor, J. F. J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S. M., Hawkins, W. T., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., & Gabriel, I. (2021). Ethical and social risks of harm from Language Models. *ArXiv, abs/2112.0*. <https://doi.org/https://doi.org/10.48550/arXiv.2112.04359>

Whitby, B. (2011). On Computable Morality An Examination of Machines as Moral Advisors. En M. Anderson y S. L. Anderson (eds.), *Machine Ethics* (pp. 138-150). Cambridge University Press.

Winter, R., & Salter, A. (2020). DeepFakes: uncovering hardcore open source on GitHub. *Pom Studies*, 7(4), 382-397. <https://doi.org/https://doi.org/10.1080/23268743.2019.1642794>

Xiao, Y., & Wang, W. Y. (2021). On Hallucination and Predictive Uncertainty in Conditional Language Generation. *ArXiv, abs/2103.1*.

Yan, L., Sha, L., Zhao, L., Li, Y.-S., Maldonado, R. M., Chen, G., Li, X., Jin, Y., & Gavsević, D. (2023). Practical and Ethical Challenges of Large Language Models in Education: A Systematic Literature Review. *ArXiv, abs/2303.1*. <https://doi.org/https://doi.org/10.48550/arXiv.2303.13379>

Yapo, A., & Weiss, J. (2018). Ethical Implications of Bias in Machine Learning. *Proceedings of the Annual Hawaii International Conference on System Sciences*. <https://doi.org/https://doi.org/10.24251/hicss.2018.668>

Yeung, K. (2019). *Responsibility and AI*. Council of Europe. <https://rm.coe.int/responsability-and-ai-en/168097d9c5>

Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, 1(6), 831-833. <https://doi.org/https://doi.org/10.1016/j.fmre.2021.11.011>

Jurisprudencia, normas y otros documentos legales

Ley 2213, Por medio de la cual se establece la vigencia permanente del decreto legislativo 806 de 2020 y se adoptan medidas para implementar las tecnologías de la información y las comunicaciones en las actuaciones judiciales, agilizar los procesos judiciales y flexibilizar la atención a los usuarios del servicio de justicia y

DESCRIPCIÓN DE
LOS RIESGOS Y
DESAFÍOS PARA
LA INTEGRIDAD
ACADÉMICA DE
APLICACIONES
GENERATIVAS DE
INTELIGENCIA
ARTIFICIAL

RISKS AND
CHALLENGES POSED
BY ARTIFICIAL
INTELLIGENCE
GENERATIVE
APPLICATIONS
FOR ACADEMIC
INTEGRITY

se dictan otras disposiciones (Congreso de la República [Colombia], 13 de junio de 2022).

Carlos Antonio Urquieta Salazar c/ Margarita Leonor Cid Lizondo, Sentencia Rol 2595-2009 (Octavo Juzgado de Garantía de Santiago [Chile], 27 de mayo de 2011). <https://juris.pjud.cl/busqueda/u?q7av>

Salvador Espitia Chávez c/ Salud Total E.P.S, Sentencia N.º 32 (Juzgado 1.º Laboral del Circuito de Cartagena [Colombia], 30 de enero de 2023).

Recibido: 20/03/2023

Aprobado: 20/06/2023