Entre la Musa y el Código: Explorando los límites del plagio y la inspiración en la era de la IA

CLARIDADES
REVISTA DE FILOSOFÍA
ISSN 1889-6855 · ISSN-E 1989-3787

Between the Muse and the Code: Exploring the boundaries of plagiarism and inspiration in the age of AI

Sara Blanco Peña

Universidad de Tübingen (Alemania)

Fecha de envío: 31/03/2024

Fecha de aceptación: 01/10/2024

DOI: 10.24310/crf.16.2.2024.19635

Resumen

En este capítulo, trato algunas de las problemáticas que surgen con la implantación de inteligencia artificial generativa (IAG) y, en concreto, cuando el remplazo de profesionales humanos por esta tecnología puede considerarse inadecuado. Considero que lo último será el caso cuando el trabajo de profesionales humanos se les enajena para convertirse en meros datos con los que entrenar los modelos de IAG remplazarán a esos mismos profesionales. La base de mi argumento reside no tanto en

el *output* de este tipo de modelos, sino en la relación de esos *outputs* tanto con los usuarios como con los profesionales humanos que hayan contribuido directa o indirectamente a la generación de los mismos.

Palabras claves

Inteligencia artificial; inteligencia artificial generativa, plagio; tecnología.

Abstract

In this chapter, I address some of the issues that arise with the implementation of generative artificial intelligence (GAI), particularly when the replacement of human professionals by this technology may be deemed inappropriate. I argue that this situation arises when the work of human professionals is alienated to become mere data used to train GAI models that, in turn, replace those same professionals. The foundation of my argument lies not so much in the *output* of these models, but rather in the relationship of these *outputs* with both the users and the human professionals who have contributed directly or indirectly to their generation.

Keywords

Artificial intelligence; generative artificial intelligence; plagiarism; technology.

I. Introducción

La inteligencia artificial (IA) es una nueva tecnología que ha llegado para quedarse. Revolucionando numerosos campos como la medicina, el derecho, la selección de personal, el diseño gráfico o la publicidad, en este artículo me gustaría centrarme en un tipo concreto de IA: la IA generativa (IAG), definida en términos generales como el conjunto de sistemas capaces de generar contenido como texto, imágenes o vídeos.

En la actualidad, no se cuenta con una definición canónica de IA y, en ocasiones, se usa dicha denominación como un término paraguas que abarca multitud de nociones complejas como *machine learning, deep learning* o *neural networks* (NNs), entre otras. Aquí voy a centrarme en estas últimas,

pues las NNs son una de las ramas de la IA más populares y prometedoras a día de hoy, y en especial en cuanto a IAG se refiere. Una NNs, como su nombre en inglés indica, es una red de nodos denominados «neuronas» capaz de procesar grandes volúmenes de información. Dichos nodos están organizados en capas, de forma que un *input* pasa por diferentes capas, activando algunos de los nodos. Las activaciones en una capa determinan las activaciones de la capa siguiente, hasta los datos pasan por la última capa y se genera un *output*. Las NNs tienen una gran variedad de aplicaciones y ya se utilizan para asistir a los humanos en diversas tareas que incluyen detección y clasificación de patrones, entre otros.

En el caso de las NNs aplicadas a la IAG, el sistema se alimenta de grandes volúmenes del tipo de información que se quiera generar (predominantemente imágenes o texto). El sistema analiza esta información y la descompone en unidades fundamentales, como palabras o píxeles. Tras este análisis, hay un proceso (más o menos) automático de predicción sobre el funcionamiento de dichas unidades. En otras palabras, la NNs «aprende» qué palabras o píxeles aparecen juntos con qué frecuencia, y cómo ordenar dichas unidades de tal modo que se produzcan outputs coherentes con los *inputs* con los que el sistema ha sido entrenado. Así, cuando al sistema se le pide que genere un cierto *output*, el modelo predice qué secuencia de unidades mínimas es más probable que correspondan con la respuesta que el usuario busca. En el caso de generadores de texto como chatGPT, el sistema es capaz de generar textos aparentemente coherentes que parecen corresponder al estilo de un ser humano que domina la lengua requerida.

La IAG es muy prometedora y abre un amplio abanico de posibilidades en la automatización de campos diversos, desde el diseño gráfico hasta la redacción de textos. En industrias como la publicidad, que demandan un gran volumen de personal y operan a ritmos vertiginosos, la IA supone una ventaja competitiva innegable. Sin embargo, esta nueva tecnología no viene sin inconvenientes. Uno de los grandes problemas éticos de la IA y, en particular de la IAG, es el uso de fuentes. En muchos casos, las compañías responsables por el desarrollo de estos sistemas se mantienen opacas respecto al origen de sus bases de datos. Es decir, se niegan a revelar de dónde salen los datos que usan para entrenar a sus NNs. Ante este desconocimiento, la línea entre entrenamiento y plagio se vuelve borrosa. Destaca el caso de las IA generadoras de imágenes, en las cuales es posible incluso demandar

imágenes que emulen el estilo de artistas particulares, siempre y cuando estos artistas sean lo suficientemente populares como para ser reconocidos por la IA. El dilema ético es claro: lejos de hacernos la vida más fácil, las IAs están empezando a ser usadas para remplazar personas en tareas ampliamente consideradas como contribuyentes al desarrollo personal, como por ejemplo, el ejercicio artístico.

La crítica anterior es ampliamente compartida por usuarios y ciertos círculos académicos. Generalmente, esta crítica se dirige a toda IAG: ya sea de textos, imágenes o vídeo. Mi propósito en este artículo es hacer una criba, a fin de señalar el problema de manera más precisa. Mi tesis es que no toda la IAG supone un potencial problema de plagio, independientemente de lo opacas que sean sus fuentes (aunque la transparencia será siempre preferible, si es que es posible). La base de mi argumento se basa en el propósito de la IAG. En contextos donde la IA se emplea para desentrañar o interpretar información compleja, la transparencia sobre las fuentes originales de dicha información puede ser menos crítica. Esto se debe a que el valor primordial aquí reside en la capacidad de la IA para facilitar la comprensión, más que en la creación de contenidos originales *per se*. Por otro lado, cuando la IAG se orienta hacia la creación de nuevo contenido, como ensayos o imágenes publicitarias, el reconocimiento y la valoración del trabajo original en el que se basa se vuelven considerablemente más importantes. En estos casos, el acto de «tomar prestado» de fuentes originales para entrenar la IA adquiere una dimensión ética que no puede ser ignorada. En resumen, las implicaciones de ser transparente acerca de las fuentes utilizadas para entrenar la IAG son variables y dependen en gran medida del propósito final de la tecnología. Mientras que en algunos casos la procedencia de la información puede ser secundaria y su transparencia opcional, en otros, especialmente aquellos con objetivos comerciales, el respeto por la autoría original es fundamental para mantener la integridad ética del proceso.

II. Problemas de la IAG: *el remplazo inadecuado*

II. 1. Conceptos preliminares: qué es la IAG

Se considera IAG a aquellos sistemas capaces de generar contenido como texto, imágenes o vídeos a través de modelos generativos. Estos modelos son entrenados con una gran cantidad de muestras del tipo de información que se

desee reproducir, para que puedan «aprender» de esas muestras e imitarlas. Así, un sistema de IAG acaba creando nuevo contenido que comparte similitudes con su base de datos de entrenamiento, a menudo en reacción a consignas o comandos específicos (Chen *et al.*, 2023: 7033). En el caso de generación de texto, el ejemplo por excelencia es el ya mencionado ChatGPT, desarrollado por OpenAI. Cuando se trata de imágenes, algunos ejemplos incluyen los modelos de DALL-E, Stable Diffusion o Midjourney (desarrollados por OpenAI, Runway y LMU, y una iniciativa de investigación independiente, respectivamente). En este tipo de sistemas, el usuario introduce un comando en formato de texto y el sistema genera una imagen a partir de dicha orden. Los sistemas de IAG son interactivos y receptivos a sugerencias adicionales, por lo que si el usuario no está satisfecho con el *output* original, es posible solicitar modificaciones (ver figura 1).

En principio, la irrupción de la IAG aparece como una innovación prometedora que ofrece resultados asombrosos generados de forma eficiente; lo que un artista humano tardaría días en producir, puede ser recreado en cuestión de segundos con modelos como DALL-E (algo más, dependiendo de la exigencia y el perfeccionismo del usuario). Con este tipo de tecnologías se ahorra tiempo, esfuerzo y dinero. Entonces, ¿cuál es el problema?

II. ii. Problemas éticos

Para determinar si una actividad es ética o no, ha de tenerse en cuenta no sólo el producto de dicha actividad sino el proceso usado para desarrollar el producto. En otras palabras, importa no sólo el qué se hace, sino también el cómo. Cuando se trata de actividades creativas, un problema común es el plagio: es decir, el hecho de hacer pasar como propia la obra ajena, total o parcialmente[1]. En algunos campos, especialmente en los artísticos, la línea entre el plagio y la inspiración es fina. En ocasiones, resulta difícil distinguir cuándo un concepto ajeno sirve como base para desarrollar uno propio, y cuándo la superposición entre ambos es tan grande qué más qué «basarse» en el concepto original, el autor se está atribuyendo ilícitamente el mérito de dicho concepto. Con la aparición de las IAGs, esta distinción se vuelve aún más complicada. Veamos con más detalle el proceso que sigue un sistema IAG para generar un *output*.

1. Para una distinción más específica entre distintos tipos de plagio, véase Martin (1994: 449).

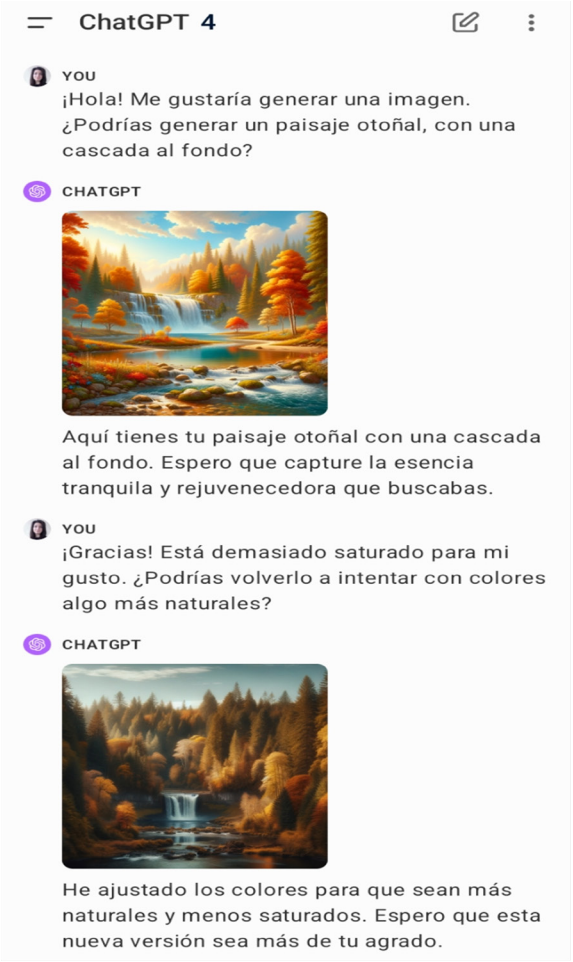

Figura 1. *Ejemplo de generación de imágenes usando la integración de DALL-E en GPT-4 (https://chat.openai.com/chat).*

Como he señalado antes, las NNs en las que se basan la mayoría de IAGs necesitan grandes bases de datos para «aprender» de forma probabilística cómo se estructuran esos datos y ser capaces de generar datos nuevos que se parezcan a aquellos con los que el sistema ha sido entrenado. Por ejemplo, para ser capaz de generar imágenes como las que aparecen en la figura 1, DALL-E ha sido entrenado con millones de imágenes de bosques, de paisajes otoñales y de cascadas (entre otros). De esta forma, el sistema es capaz de reconocer qué tipo de pixeles aparecen con qué frecuencia y en

qué tipo de configuraciones en las imágenes asociadas a dichas categorías. Así, cuando el usuario usa estas palabras en sus comandos, DALL-E reproduce dichas configuraciones: no de forma idéntica, sino que genera nuevas secuencias de píxeles que se parezcan lo suficiente a las secuencias que ha ido identificando en su entrenamiento.

Así, las IAGs necesitan obras originales para ser entrenadas y para «aprender» a generar sus propias obras. En principio, esto no tendría por qué ser un problema: todo artista necesita inspiración, y a escribir se aprende leyendo. Para determinar si el aprendizaje artificial supone un problema ético o no, necesitamos prestar atención a dos cuestiones:

a) Qué tipo de *outputs* se generan, y cuánto se parecen a las obras originales con las que el sistema ha sido entrenado.

b) Cómo se han usado las obras de la base de datos de entrenamiento para llegar a producir un *output* «original[2]».

El punto *a* apunta a lo que podría ser un criterio para distinguir entre un uso ético y no ético de la IAG[3]: si es posible encontrar similitudes entre el *output* generado y los datos de entrenamiento que sean lo suficientemente evidentes, se podría decir que la IAG está plagiando las obras con las que ha sido entrenada. No resulta controvertido señalar este tipo de uso de la IA como no ético (Somepalli et al., 2023: 6049). Sin embargo, cabe señalar que, intuitivamente, el plagio es una actividad exclusivamente humana, ya que requiere de una cierta intención fraudulenta, de forma similar a otras actitudes reprochables como la manipulación (para un argumento detallado sobre la segunda, véase Klenk (2022)). En este artículo, me tomo la libertad de usar el término «plagio» para referirme a IAG, haciendo referencia al fenómeno que Somepalli et al. denominan réplica de contenidos («*content replication*»). En este sentido, el plagio se refiere a la reproducción de parte

2. Entre comillas, porque si es original o no, está por ver. Más detalles sobre la cuestión de la originalidad y la creatividad en sistemas de IAG puede encontrarse en Eshraghian (2020).

3. Este criterio es análogo a los criterios usados para detectar plagio por parte de autores humanos. En el caso de la creación de textos, softwares como Turnitin discriminan entre textos plagiados y originales en base al porcentaje de texto que presente coincidencias exactas con texto de terceras partes. Es decir, se determina si algo es plagio o no en base a un criterio cuantitativo: cuando el porcentaje de coincidencias encontradas supera cierto número, se considera que la obra a evaluar no es original. Este es un criterio ampliamente aceptado en el mundo académico, estando en vigor en universidades de todo el mundo para evaluar la originalidad de la obra de sus estudiantes.

o la totalidad de la base de datos de entrenamiento en el *output* generado. Por ejemplo, consideremos un sistema de IAG entrenado con fotografías de personas reales, como el generador de rostros This Person Does Not Exist[4]. Si se le pide generar un retrato nuevo y el resultado reproduce detalles específicos y reconocibles de una persona en particular, esto podría ser visto como plagio. Esto sucede porque el sistema no está creando un rostro nuevo de alguien que no existe, sino replicando los rasgos de una persona concreta que aprendió durante su entrenamiento.

Por otro lado, el punto *b* se centra no tanto en el *output* generado sino en el proceso de generación de dicho *output*. En concreto, me refiero aquí a cuestiones como: ¿Cuál es el origen de la base de datos de entrenamiento? ¿Es la base de datos de entrenamiento apropiada para el tipo de *output* que se pretende generar? ¿Es legítimo el uso que se está haciendo de las obras que conforman la base de entrenamiento?, etc. Este tipo de preguntas señalan una variedad de problemas que se dan con el uso actual de la IAG y con la opacidad de sus fuentes. La opacidad de la mayoría[5] de desarrolladores respecto a sus datos de entrenamiento puede encubrir problemas como el uso de datos obtenidos de forma fraudulenta, de bases de datos sesgadas, de uso ilegítimo de datos, etc. Me interesa en concreto un problema que atañe de manera especial al fenómeno de la IAG, y cuya crítica es particularmente acuciante en lo que a IAs generadoras de imágenes se refiere. Me refiero al uso de IAG para generar productos que, cuando son obra humana, suponen una realización personal del autor. Es decir, sustituir ejercicios que, cuando son ejercidos por personas, llevan a la realización de las mismas, por sus imitaciones artificiales. Me refiero a este fenómeno como *remplazo inadecuado*[6].

El tipo de cuestiones enumeradas arriba señalan problemas graves que pueden darse con la implementación de IA(G) en sociedad. Sin embargo, muchas de estas preguntas esbozan dilemas éticos aun sin planteamiento ni respuestas claras, y es tentador buscar la forma de llevar a cabo la implementación de IA(G) de forma responsable. Nos encontramos en un punto del desarrollo de la IA en el que parece que gran parte de nosotros

4. Véase https://this-person-does-not-exist.com
5. En el caso de modelos como Stable Difussion, se conoce la base de datos que se usa para entrenar al modelo, pero autores particulares no tienen opción a que sus obras no formen parte de dicha base de datos. Analizo este caso en más detalle en la siguiente sección.
6. En la siguiente sección, profundizo en esta idea.

quisiéramos tener nuestra tarta artificial y comérnosla: quiero ser capaz de replicar lenguaje natural humano para agilizar tareas arduas y tediosas, pero no quiero que esa réplica desemboque en un escenario distópico en el que se reproduzcan de manera sistemática e incontrolable sesgos raciales, de género[7], etc. En este artículo, no prometo una solución a este problema. Lo que sí ofrezco es el primer paso hacia la misma: un análisis de ese remplazo inadecuado del ejercicio humano por aprendizaje artificial. Es decir, hacer explícita la problemática de algunos usos concretos de las IAGs, clarificando cuándo remplazar personas por IAG resulta inadecuado y por qué. El gran foco de este artículo está puesto en la generación artificial de imágenes. Al final del mismo, haré la comparativa con la generación artificial de textos, y con distintas aplicaciones que pueden darse de ambas actividades. Para ello, en la siguiente sección me dispongo a evaluar el papel de las personas en la generación artificial de contenido y contrastarlo con su papel como creadores humanos. A continuación, ahondaré en el concepto de plagio en lo que a IAG se refiere y en las consecuencias de sistemas de IA accediendo a contenido generado por personas. Finalmente, propondré algunos usos de IAG menos conflictivos y elucidaré la diferencia con otros usos éticamente problemáticos.

III. Análisis del remplazo inadecuado

III. i. Nuevas formas de alienación

Desde hace décadas, investigadores en informática y computación trabajan en el desarrollo de IA, bajo la promesa de una nueva tecnología capaz de realizar tareas que hasta el momento requerían de intelecto humano. Ya en 1950, Turing (1950) planteaba la pregunta de si pueden pensar las máquinas. Para responderla, introdujo el criterio de que, si un humano no es capaz de distinguir si está interactuando con una máquina o con otro ser humano, la máquina se considera pensante[8]. En los últimos años, con la aparición de

7. Más detalles sobre cómo generadores artificiales de texto como chatGPT reproducen sesgos de género pueden ser encontrados en Gross (2023).
8. Este criterio queda recogido en lo que se conoce como Test de Turing, denominado como Imitation Game en el artículo citado arriba.

deepfakes[9] prácticamente indistinguibles de sus homólogos originales, este horizonte se presenta más que tangible. A día de hoy, la meta de la IA va más allá de la imitación que Turing vislumbraba hace más de medio siglo: la aspiración ya no es sólo desarrollar sistemas que realicen tareas como si fueran humanos, sino que nos superen en eficiencia y precisión. Una prueba fehaciente de ello es la normalización de asistentes virtuales y *chatbots*[10], que ya han sustituido a lo que antes eran operadores humanos. No sólo eso, sino que, hoy, el problema no es tanto la calidad de la imitación que algunos de estos asistentes ofrecen, sino el hecho de que muchas compañías no hacen transparente su uso de IA, dejando poco claro para el usuario si sus interacciones tienen lugar con una persona o con uno de estos sistemas. A día de hoy, los ojos están puestos en el desarrollo de modelos que superen las capacidades humanas en ámbitos como la programación, la imagen médica o la selección de personal[11]. La IA abre un mundo de posibilidades y promete liberarnos de la ejecución de tareas arduas y tediosas, como la revisión de textos, la dispensación de fármacos o la clasificación y empaquetamiento de productos varios, por ejemplo. Sin embargo, la sombra de esta promesa es alargada, ya que la IA es capaz de ejecutar no sólo este tipo de tareas, sino muchas otras (Stokel-Walker and Van Noorden, 2023).

Como he venido anticipando, los sistemas de IAG son capaces de desempeñar tareas creativas de forma mucho más eficiente que profesionales humanos. En los últimos meses, las imágenes generadas con IAG pueblan con abundancia no sólo las redes sociales, sino campañas publicitarias[12],

9. Un *deepfake* es una imagen, vídeo o audio generado con IAG que imita la apariencia y/o sonido de una persona. Los deepfakes se caracterizan por ser contenidos altamente realistas, pese a ser ficticios, siendo muy difícil de distinguir de una imagen, vídeo o audio real de la persona cuya apariencia se está suplantando. Sirva como ejemplo son los vídeos virales en los que usuarios de YouTube insertaron el rostro de Nicolas Cage en el de otros actores y actrices, haciendo aparecer al actor en películas en las que nunca participó (Wagner & Blewer, 2019: p. 37).

10. Ejemplos de estas aplicaciones son asistentes como Siri, Alexa, o la nacional —y ya extinta— Irene de Renfe.

11. Sirven como ejemplos en los campos mencionados modelos como GitHub Copilot, IBM Watson for Oncology o Entelo, respectivamente.

12. Un ejemplo reciente es el catálogo de Navidad de la empresa Juguettos: https://catalogo.juguettos.com/.navidad/#page=1

cubiertas de libros[13], e incluso campañas gubernamentales[14]. En estos casos, la IAG no parece estarse usando para ejecutar tareas tediosas o repetitivas, sino trabajos artísticos que suelen considerarse vocacionales, tales como el diseño gráfico y la creación artística. ¿Pero por qué se ha acabado usando IAG para este tipo de tareas? ¿No era la gran promesa liberarnos de aquello que nos resulta tedioso? La respuesta es simple, pero no por ello menos decepcionante. La IAG es una tecnología que permite realizar tareas antes efectuadas por personas de forma mucho más eficiente, ya que estos modelos ahorran tiempo y dinero. De acuerdo a los datos de Sotille (2023), pagar a un artista el salario mínimo por una hora de trabajo es 90 veces más caro que el generador de arte más barato. Por tanto, no resulta sorprendente que en nuestra sociedad capitalista prime el beneficio capital frente a un potencial aumento de la calidad de vida a la hora de implementar esta tecnología. La meta última no es hacer la vida del ciudadano medio más fácil, sino obtener de él el mayor beneficio posible, vendiéndole productos que han sido generados a un coste mucho menor de lo que habría costado emplear a un equipo de profesionales. Se introduce así un matiz clave: la IA tiene el potencial de facilitar la vida de las personas, pero ahora la cuestión es la vida de qué personas. Para sorpresa de nadie, la respuesta es la vida de los propietarios de los sistemas de IA.

Pero me estoy adelantando. Recapitulemos y repasemos el proceso de generación de una imagen mediante IAG con fines comerciales. Como ya sabemos, un modelo IAG necesita primero ser entrenado usando una base de datos. Después, normalmente hay un periodo de corrección manual de lo que ha «aprendido» el modelo. A continuación, se sigue entrenando el modelo sometiéndolo a varias rondas de exposición a datos adicionales y corrección manual. Cuando los *outputs* generados son satisfactorios, se

13. Un ejemplo claro es la cubierta de tapa blanda para Reino Unido de House of Earth and Blood, un bestseller de Sarah J. Maas. Más información sobre el proceso y controversia puede encontrarse aquí: https://www.theverge.com/2023/5/15/23724102/sarah-j-maas-ai-generated-book-cover-bloomsbury-house-of-earth-and-blood

14. El 11 de febrero de 2024, el Ministerio de Juventud e Infancia del Gobierno de España hizo uso de imágenes generadas por IAG para promocionar el día de la mujer y la niña en la ciencia en la red social X (antes conocida como Twitter). Tras una avalancha de críticas por parte de los usuarios de dicha red social, el Ministerio se vio obligado a retirar dichas imágenes y ofrecer una disculpa pública que puede encontrarse aquí: https://twitter.com/JuventudInfGob/status/1756705331780923814

considera que el modelo está listo para ser usado por usuarios y generar sus propios datos. Ilustro esto con un ejemplo concreto: digamos que queremos desarrollar un modelo capaz de generar paisajes. Para ello, se alimenta el modelo con un gran volumen de imágenes de paisajes, cuanto más variadas mejor. El modelo descompondrá estas imágenes en píxeles y, siguiendo un algoritmo, atribuirá a cada pixel un peso, que determina su importancia en la imagen. Esto sirve no para «entender» qué es una cascada o un árbol, sino más bien para determinar cómo de probable es que los píxeles que pertenecen a las tonalidades que normalmente se corresponden con agua aparezcan juntos, formando imágenes de cascadas. Es decir, el modelo hace una estimación probabilística de qué elementos aparecen en qué configuraciones. El siguiente paso es ser capaz de reproducir este tipo configuraciones; que no copiar los ejemplos exactos de la base de datos. En otras palabras, de generar sus propias cascadas y árboles, lo suficientemente parecidos a los de la base de datos de entrenamiento como para ser reconocidos como cascadas y árboles, pero lo suficientemente diferentes como para ser considerados «nuevos». A continuación, un equipo de personas revisa las primeras rondas de entrenamiento, y corrige las anomalías del modelo; es decir, indican al modelo que las imágenes que incluyen elementos como cascadas que no desembocan en ningún río, o árboles con raíces surgiendo de ninguna parte, no son válidas. Se sigue entrenando al modelo, hasta que genere mayoritariamente paisajes aceptables y se considere listo para usar por el usuario.

Continuando con el ejemplo anterior, digamos que una empresa decide usarlo para generar imágenes publicitarias, ahorrándose los sueldos de ilustradores o fotógrafos, diseñadores gráficos, etc. Sin embargo, la labor de estos profesionales sigue siendo necesaria: ya no para elaborar el producto final, sino para producir los paisajes con los que entrenar a la IAG que genera el producto final. Recordemos que el modelo necesita infinidad de imágenes para ser entrenado. ¿De dónde salen estas imágenes? En su mayoría de esos profesionales de los que sale más barato prescindir, pues la IAG permite acceder a su obra de forma indirecta y gratuita. Por poner un caso más concreto: el modelo de IAG Stable Diffusion utiliza una red masiva de arte digital extraído de Internet, de una base de datos llamada LAION-5B (Schuhmann et al., 2022). Los artistas cuyas obras conforman LAION-5B no pueden optar por que sus obras no sean incluidas en dicha

base y no se usen para el entrenamiento de Stable Diffusion u otros modelos de IAG. Esto preocupa a algunos artistas, que afirman que Stable Diffusion se basa en sus obras para crear sus propias imágenes, pero no se les acredita ni compensa por su trabajo (Metz, 2022). Es decir, en lugar de retribuir a profesionales por un trabajo que les llena, se sigue usando su trabajo de forma ilegítima y sin retribución. Se constituye así una nueva forma de alienación, que convierte el trabajo creativo como expresión genuina de humanidad en una enajenación de la que no se llega a ver nunca el fruto, y cuyo beneficio se llevan terceras partes. En este trabajo, empleo el término «alienación» en un sentido que resuena profundamente con la crítica marxista de la estructura laboral capitalista. En sus *Manuscritos Económico-Filosóficos*, Marx (2015) describe la alienación como el proceso por el cual los trabajadores se vuelven extraños frente a los productos de su labor. Esta alienación no solo ocurre porque los trabajadores no poseen los productos que crean, sino porque el acto de producción mismo, bajo condiciones capitalistas, se transforma en una actividad ajena y hostil a su propia naturaleza. Siguiendo a Marx, el producto del trabajo del trabajador es extraño al propio trabajador, y cuánto más trabaja más alimenta lo que le es extraño (Marx, 2015: 137). Aplicado al tema que nos ocupa, el paralelo puede trazarse con los creadores y artistas se encuentran alienados respecto a sus obras frente a las obras generadas por IAGs. La introducción de esta tecnología convierte a las obras originales producidas por humanos en contenido para entrenar IAGs, en lugar de un producto final en sí mismo. Así, se genera una distancia entre lo que las personas producen y el producto final que constituyen lo que las IAGs acaban generando, en la mayoría de los casos, de forma impredecible.

De ahí que este remplazo de profesionales humanos por IAG sea inadecuado: no se remplazan personas en ámbitos que beneficien a su calidad de vida, sino que, por el contrario, se les enajena de actividades que contribuyen al desarrollo de su humanidad. La promesa original de la IA era hacer nuestra vida más fácil, aligerando trabajos que se desempeñan por ser necesarios para un fin mayor, pero no enriquecedores de por sí. En algunos casos, nuevas tecnologías ya están siendo usadas para estos fines, evitando que trabajadores humanos asuman riesgos innecesarios. Algunos ejemplos son la utilización de drones para la inspección de infraestructura, robots desactivadores de bombas, en la industria petrolífera o en minería

subterránea. Sin embargo, sigue habiendo muchas tareas no especializadas que, pese a poder ser realizadas, o al menos asistidas, por IA, siguen siendo desempeñadas por personas cobrando sueldos ínfimos. Este panorama evidencia una realidad incómoda: la implementación de la IAG no siempre busca priorizar la mejora de la calidad de vida humana o aliviar las cargas de trabajo. Más bien, parece que la verdadera intención detrás de muchos desarrollos en IAG apunta hacia la maximización de beneficios y la rentabilidad de las industrias, incluso si ello significa perpetuar o agravar las condiciones de explotación humana. Es el caso de los artistas cuyas obras se usan de forma ilícita para constituir bases de datos de entrenamiento, así como de autores de otros tipos de contenido cuyas obras se usen de manera similar.

III. ii. Del plagio a la inspiración

La pregunta ahora es: ¿es IAG sinónimo de alienación? ¿no es posible reconciliar la eficiencia que esta tecnología ofrece con el respeto hacia el trabajo de profesionales humanos? En principio, la respuesta a esta primera pregunta no tiene por qué ser afirmativa. La crítica que he desarrollado en los párrafos superiores se centra no en la IAG en sí, sino en el proceso ilícito de recopilación y utilización de datos para crear bases de entrenamiento, donde se acumulan imágenes (u otros contenidos) obtenidas de manera cuestionable; es decir, sin la debida remuneración ni permiso de utilización de los autores. Por tanto, se podría concluir que, siempre y cuando este no sea el caso y el sistema IAG se entrene respetando la autoría de los datos que se usen en su entrenamiento, la puerta está abierta hacia posibles usos éticos.

En concreto, mi crítica hacia la IAG no se dirige hacia el proceso de generación artificial de contenido per se, sino a lo que se hace con ese contenido en contraste con el papel de las personas que indirectamente han contribuido a su generación. En otras palabras, considero que el remplazo de profesionales por IAG es inadecuado cuando el trabajo de estos se enajena para entrenar a dicha tecnología. Así, la IAG puede convertirse en una nueva forma de alienación cuando el trabajo de profesionales humanos se enajena de su autoría o derechos, utilizándolo para entrenar sistemas de IA que luego producirán beneficios para otros, sin reconocer debidamente la contribución de dichos profesionales.

Por otro lado, cuando los derechos de los profesionales humanos se respetan, la IAG puede ser una gran herramienta para esos mismos profesionales. Por ejemplo, herramientas como la integración de Adobe Firefly[15] en la última versión de Photoshop, permiten a artistas ampliar los fondos de sus propias imágenes para seguir trabajando en ellas. Así, podría decirse que el uso de IAG no tiene por qué necesariamente caer en el plagio, sino que puede servir como inspiración; algo así como una musa digital, ofreciendo un punto de partida sobre el cual los profesionales pueden construir y refinarse. Esto se da cuando el producto generado por IAG no es el producto final, sino una base sobre la que seguir trabajando de forma más eficiente. Encontramos ejemplos en el caso de la generación de imágenes, donde la IAG puede proporcionar referencias visuales iniciales, o en la creación de textos, donde escritores la utilizan para afinar aspectos gramaticales de sus obras o superar el bloqueo creativo mediante sugerencias que estimulan nuevas ideas. De este modo, la IAG puede usarse para enriquecer la creatividad humana, y no necesariamente para alienar a los autores de su propio trabajo creativo.

III. iii. Distinción fundamental

En las secciones anteriores he ido perfilando un argumento que me lleva a discernir entre dos maneras de entender la IAG: una como medio de enajenación y otra como herramienta genuinamente enriquecedora. Esta distinción fundamental radica en cómo los creadores humanos interactúan y se benefician del producto de la IAG. Si la tecnología actúa como un catalizador de información que contribuye a la autorrealización de los individuos, entonces estamos ante un uso de la IAG que enriquece la experiencia humana. Si, por el contrario, se explota el trabajo de profesionales

15. Más información sobre Adobe Firefly puede ser encontrada aquí: https://www.adobe.com/es/products/firefly.html. De acuerdo con su apartado de preguntas frecuentes, Adobe está «entrenando [su] modelo inicial comercial de Firefly con contenido con licencia, como Adobe Stock, y contenidos de dominio público cuyos derechos de autor han caducado. Además, como socio fundador de la Iniciativa de Autenticidad del Contenido (Content Authenticity Initiative, CAI) Adobe establece el estándar del sector de la IA generativa responsable». Aunque aún es incierto si este criterio es suficiente para considerar un modelo de IAG «responsable», al menos es un buen primer paso hacia la consideración de los derechos de los autores humanos que producen los datos con los que estos modelos se entrenan.

humanos sólo para alimentar tecnologías que les quitan reconocimiento y beneficios, estamos ante un dilema ético importante. Entender esta distinción es crucial para comprender qué usos de la IAG son problemáticos y cómo asegurarnos de que su implementación en sociedad fomente el desarrollo humano, en lugar de sustituirlo.

Hasta ahora, he puesto el foco en la generación de imágenes, ya que modelos como Stable Diffusion o DALL-E han revolucionado la industria, aunque la recepción por parte del público general siga siendo objeto de debate (Bran et al., 2023). Sin embargo, la generación de imágenes no es el único tipo de IAG que genera controversia. Los generadores de texto, como ChatGPT, han alcanzado una popularidad inmensa, transformándose en una herramienta cotidiana para una amplia gama de usuarios. A pesar de su éxito, no están exentos de críticas, aunque estas difieren sustancialmente de las dirigidas a la generación de imágenes por IAG.

En el ámbito de los textos, las preocupaciones se centran menos en el plagio directo de autores individuales y más en la falta de originalidad por parte de quienes utilizan esta tecnología. Dado que el lenguaje no es un medio visual, resulta complicado, si no imposible, identificar el «estilo» específico que inspira a los Modelos de Lenguaje de Gran Tamaño (LLMs) para generar contenido nuevo.

Sin embargo, existen inquietudes legítimas respecto a la autoría y la propiedad intelectual en este tipo de IAG. En contraste con el caso de la generación de imágenes, donde artistas individuales se ven afectados, en el caso de los textos generados por IA, las entidades más perjudicadas tienden a ser instituciones, como periódicos. Un ejemplo notable es la denuncia planteada por el New York Times a OpenAI y Microsoft[16], que refleja esta preocupación a nivel corporativo. De acuerdo con el periódico, sus artículos habrían sido usados para entrenar la NNs de ChatGPT, con la que ahora compiten en cuanto a la búsqueda de información. No sólo se trata de que el modelo de OpenAI ofrezca la misma información[17] de forma más accesible (pues se adapta de forma interactiva a las preguntas concretas del usuario), sino que ChatGPT ofrece dicha información de forma gratuita,

16. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

17. En cuanto a contenido, ya que ha usado los artículos del New York Times como entrenamiento (entre otros).

mientras que el New York Times, como muchos otros periódicos, protege sus textos tras un muro de pago. Este escenario plantea una problemática distinta a la del «remplazo inadecuado». La principal diferencia es que aquí los afectados no son autores individuales, sino instituciones. Los periódicos y grandes medios, al ser entidades corporativas, no pueden ser «alienados» de su trabajo de la misma manera que un trabajador individual. Por lo tanto, el problema que surge con este uso de la IAG no es tanto el plagio, sino el acceso a la información. Esto se complica porque, como en el caso del New York Times, la disputa se basa en una presunta violación de derechos de autor, donde el «autor» no son las personas que redactan los artículos, sino el periódico como institución. Aunque este análisis podría extenderse a incluir aspectos de alienación observados en ciertas prácticas corporativas, esa es una consideración que merece una exploración detallada por sí misma[18].

Con esto, no quiero decir que el remplazo inadecuado aplique sólo a la generación de imágenes, mientras que en el caso de los textos la enajenación del trabajo de autores individuales no sea posible. Del mismo modo que ilustradores, fotógrafos y diseñadores gráficos pueden verse alienados de su trabajo cuando este se usa de forma ilícita en el entrenamiento de IAG, lo mismo puede suceder con escritores y redactores, aunque en el caso de estos últimos reconocer qué textos han sido usados para generar *outputs* sea más difícil. La distinción que propongo en este artículo no tiene que ver con el tipo de *output* que se genera, bien texto o bien imágenes; tiene que ver con la relación entre ese *output*, los usuarios y las personas que han contribuido a él, directa o indirectamente. Cuando los afectados son grandes instituciones, que por sí mismas ya alienaban de su propio trabajo a sus trabajadores, no se da un remplazo inadecuado. Irónicamente, puede darse un fenómeno casi contrario: sistemas de IAG acceden al trabajo que previamente había sido

18. Un análisis más detallado de cómo aplicar el concepto de alienación marxista al mercado capitalista contemporáneo va más allá del alcance de este artículo. Pueden encontrarse propuestas con dicho fin en Choquet (2013) y en Øversveen (2021), con reinterpretaciones del término partiendo de su uso en los Manuscritos Económico-Filosóficos y escritos más tardíos, respectivamente. Este artículo comparte punto de partida con el análisis de Choquet, con cuyas conclusiones coincido. De acuerdo con él, «[..] la alienación del trabajo humano institucionalizada coercitivamente a través del sistema salarial capitalista ha sido (y *sigue siendo*) el catalizador más eficaz de la alienación social (lato sensu) en la sociedad contemporánea» (Choquet, 2021, p. 107, la traducción y el énfasis son míos).

alienado, lo procesan y lo devuelven al público, de forma que un mayor número de personas puedan enriquecerse de un producto que ya había sido enajenado de sus autores[19]. Por este motivo, y pese a la controversia legal, considero este uso de la IAG como un caso de herramienta enriquecedora, ya que contribuye a la democratización de la información.

IV. Hacia una colaboración enriquecedora

En las secciones anteriores, he desarrollado el grueso de mi argumento. En esta última, me dispongo a ampliar mi visión sobre el uso de IAG como herramienta legítima y a sugerir opciones para establecer una colaboración enriquecedora entre este tipo de modelos y creadores humanos.

Hasta ahora, he usado términos como «enriquecimiento» o «autorrealización» de forma más o menos genérica. Siendo más concreta, con «enriquecimiento» me refiero al proceso dinámico de transformación personal que va más allá de lo material, enfocándose en la plenitud de la vida humana, la libertad, la creatividad, y la integración con la comunidad y el entorno. Bajo mi propuesta, la IAG puede suponer una herramienta enriquecedora cuando la relación entre esta y los creadores humanos contribuye este tipo de autorrealización, tanto de esos mismos creadores como de los usuarios (en el caso de no ser los mismos). Esto puede darse cuando los *outputs* generados por la IAG potencian la creatividad de sus usuarios, incentivan su curiosidad o facilitan la comprensión del mundo que les rodea. Por ejemplo, cuando modelos como ChatGPT se usan para para hacer la información más accesible —traduciéndola, resumiéndola o simplificándola—, su valor como herramienta para la autorrealización humana se hace evidente. En estos casos, la IAG no sólo democratiza el acceso a la información, sino que también potencia la capacidad de los individuos para interactuar con el conocimiento de manera más profunda

19. Reconozco que este punto es controvertido. Se podría argumentar que los redactores son remunerados por los periódicos, y que en caso de que el periódico no tenga suficientes suscriptores (porque esos suscriptores prefieren acceder a la información usando ChatGPT), el fracaso del periódico repercute en los redactores, que podrían perder su trabajo. Sin embargo, mi planteamiento es más general. Grandes periódicos como el New York Times tendrían que enfrentarse a una fluctuación muy significativa de suscriptores para que su plantilla se viese afectada. Cabe destacar que, en el caso de periódicos e instituciones más pequeñas, la situación es muy distinta. Pero irónicamente, las denuncias no vienen por parte de pequeñas instituciones.

y personalizada, subrayando su papel como una auténtica musa digital al servicio del progreso humano.

Así, mi distinción se basa en la relación entre la tecnología, las personas y el mundo. Aunque la aplicación a los escenarios que surgen con la irrupción de la IAG es novedosa, el trasfondo del argumento dista de ser nuevo. Mientras que, con frecuencia, en filosofía de la tecnología esta última puede considerarse como un medio para conseguir un fin, no siempre tiene por qué ser así. Un ejemplo de lo contrario reside en *La Pregunta por la Técnica* de Heidegger. Heidegger propuso entender la tecnología como algo más, pudiendo nuestra relación con ella ir más allá de la explotación de los recursos naturales para generar productos que sirvan a nuestras necesidades. En sus propias palabras, «[…] la tecnología no es un mero medio. La tecnología es una forma de revelación» (Heidegger, 1977: 12)[20]. ¿Pero qué es lo que la tecnología tiene la capacidad de revelar? Con «revelación», Heidegger se refiere a su idea del Ser y a la búsqueda de una especie de autenticidad subyacente a todas las cosas. Las pretensiones de mi argumento son más limitadas. Sin embargo, sí coincido con Heidegger en que a través de la tecnología es posible revelar algo de nosotros mismos, a través de nuestra creatividad y sed de conocimiento. Es a esto a lo que me refiero cuando hablo de la IAG como una herramienta enriquededora: la tecnología como *poeisis*, como alumbramiento. Alumbramiento de algo que ya estaba en nosotros, pero a lo que haciendo uso de nuevas herramientas es posible acceder de forma más eficiente.

Entendiendo la tecnología de este modo, no es la IAG en sí la que nos plantea un dilema ético, sino ciertos usos de la misma. Insistiendo en la distinción heideggeriana entre la tecnología como mero medio y la tecnología como *poesis*, el foco de mi crítica se centra la primera acepción, en el empleo de la IAG para la producción en masa. El verdadero problema surge cuando usamos IAG no para enriquecer nuestra experiencia humana, sino de una manera que nos desvincula y enajena de nuestras propias creaciones. Al hacerlo, se incide en esta nueva forma de alienación del trabajo humano, corriendo el riesgo de caer en un remplazo inadecuado de profesionales humanos por modelos de IAG.

Este último punto, tampoco es nuevo. El remplazo inadecuado se refiere a una perspectiva del trabajo y su relación con los trabajadores en términos

20. Mi traducción (de la traducción al inglés de William Lovitt).

marxistas, identificando como principal problema la alienación del fruto del trabajo propio (Marx, 2015: 114-115). El problema de la alienación viene de lejos: simplemente va adoptando nuevas formas. Sin embargo, es importante identificar correctamente el problema para encontrar soluciones adecuadas. En el caso de la IAG, el problema del remplazo inadecuado se confunde con frecuencia con un problema de plagio y autoría. Sin embargo, esto no siempre es el caso, y dependerá de cómo se traten los datos de entrenamiento de la IAG y los fines para los que se usen los *outputs* generados. Tiene más sentido hablar de plagio cuando la IAG se usa con fines comerciales. Sin embargo, cuando atañe a usuarios concretos y, sobre todo, cuando lo que se genera no es un producto sino que se dan interacciones entre humanos y IAG que contribuyen al desarrollo personal de los primeros, estamos ante un tipo de dinámica diferente.

V. Conclusión

En este artículo, he señalado algunos de los problemas éticos que la IAG plantea, identificando el remplazo inadecuado como el problema principal. Con remplazo inadecuado, me refiero al uso de sistemas de IAG en lugar de profesionales humanos para tareas que, en principio, contribuyen al desarrollo personal. El calificativo de «inadecuado», viene del hecho de que, en ocasiones, la IAG puede usarse para alienar a profesionales humanos de su trabajo al ser convertido en meras bases de datos de entrenamiento. Identificar este problema es relevante, a fin de no confundirlo con problemas clásicos de autoría y plagio. Mientras que estos problemas también pueden darse según cómo se use la IAG y cómo se obtengan los datos necesarios para su entrenamiento, considero el problema del remplazo inadecuado más relevante, sobre todo teniendo en cuenta el futuro tentativo de la IAG en nuestra sociedad.

En resumen, el problema principal que supone la IAG no es la IAG en sí, sino ciertos usos que se pueden hacer de ella. Propongo un uso constructivo de esta tecnología que lleve a una colaboración enriquecedora entre profesionales humanos y sistemas de IAG. Considero que dicha colaboración es posible, siempre y cuando los *outputs* generados por la IAG contribuyan al desarrollo personal de los individuos, sin alienar el trabajo ajeno en el proceso.

Referencias bilbiográficas

Bran, E., Rughiniş, C., Nadoleanu, G. and Flaherty, M. G. (2023). «The Emerging Social Status of Generative AI: Vocabularies of AI Competence in Public Discourse», in *24th International Conference on Control Systems and Computer Science (CSCS)*, pp. 391–398. doi:10.1109/CSCS59211.2023.00068.

Chen, C., Fu, J., & Lyu, L. (2023). «A Pathway Towards Responsible AI Generated Content», in Elkind, E. (ed.) *Proceedings of the Thirty-Second International Joint Conference onArtificial Intelligence, IJCAI-23*. International Joint Conferences on Artificial Intelligence Organization, pp. 7033–7038. doi:10.24963/ijcai.2023/803.

Choquet, P.-L. (2021). «Alienation and the Task of Geo-social Critique», *European Journal of Social Theory*. SAGE Publications Sage UK: London, England, 24(1), pp. 105–122.

Eshraghian, J. K. (2020). «Human ownership of artificial creativity», *Nature Machine Intelligence*, 2(3), pp. 157–160. doi:10.1038/s42256-020-0161-x.

Gross, N. (2023). «What chatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI», *Social Sciences*. MDPI, 12(8), p. 435.

Heidegger, M. (1977). *The Question Concerning Technology and Other Essays*. New York: Garland Publishing.

Martin, B. (1994). «Plagiarism: a Misplaced Emphasis», *Journal of Information Ethics*, 3(2), pp. 447–451.

Marx, K. (2015). *Manuscritos Económico-Filosóficos de 1844*. Ediciones Colihue SRL.

Metz, R. (2022). «These artists found out their work was used to train AI. Now they're furious». *CNN*. Available at: https://edition.cnn.com/2022/10/21/tech/artists-ai-images/index.html (Accessed: March 22, 2024).

Øversveen, E. (2021). «Capitalism and Alienation: Towards a Marxist Theory of Alienation for the 21st Century». *European Journal of Social Theory*. SAGE Publications Ltd, 25(3), pp. 440–457. doi:10.1177/13684310211021579.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., & Wortsman, M. (2022).

«Laion-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models». *Advances in Neural Information Processing Systems*, 35, pp. 25278–25294.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., & Goldstein, T. (2023). «Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models», in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058.

Sotille, Z. (2023). «What to know about Lensa, the AI portrait app all over social media». *CNN*. Available at: https://edition.cnn.com/style/article/lensa-ai-app-art-explainer-trnd/index.html (Accessed: March 22, 2024).

Stokel-Walker, C., & Van Noorden, R. (2023). «The promise and peril of generative AI». *Nature*, 614(1), pp. 214–216.

Turing, A. (1950). «Computing machinery and intelligence». *Mind*, LIX(236), pp. 433–460.

Wagner, T. L., & Blewer, A. (2019). «"The Word Real Is No Longer Real": Deepfakes, Gender, and the Challenges of AI-Altered Video». *Open Information Science*. (Open Information Science), 3(1), pp. 32–46. doi:doi:10.1515/opis-2019-0003.

Sara Blanco Peña: Doctoranda en la Universidad de Tübingen (Alemania). Actualmente se encuentra finalizando su tesis doctoral sobre la confianza y la responsabilidad moral en la inteligencia artificial (IA).

Líneas de investigación:

– Ética de la IA, epistemología en IA explicable, confianza en la tecnología.

Publicaciones recientes:

– Blanco, S. (2022). Trust and Explainable AI: Promises and Limitations. Ethicomp Conference Proceedings, pp. 245-256.

Correo-e: sara.blanco@uni-tuebingen.de