Generadores de imágenes por una IA a partir de una descripción: Casos Dall·E y Stable Diffusion

Profesión y Sociedad. Entrega 2

Grupo 14

1. Introducción

En el último año, se ha producido un rápido aumento de los modelos de aprendizaje automático capaces de convertir las descripciones de texto escritas por el usuario en imágenes naturalistas. Estas IA están disponibles para cualquier usuario y, con poco o ningún conocimiento previo, son usadas para generar cantidades ingentes de imágenes al día. La publicidad de la industria, el bombo y la facilidad de acceso ya han llevado a millones de usuarios a generar millones de imágenes al día, lo que que nos ha motivado para investigar más sobre el tema y plantearnos las cuestiones éticas alredor de estas IA generadoras.

Sin que muchos usuarios lo sepan, estos modelos se han entrenado con conjuntos de datos masivos de imágenes y texto extraídas de la web, que se sabe que contienen contenidos estereotipados, tóxicos y pornográficos. Esto motiva serias preocupaciones sobre los sesgos en estos modelos que proliferan a escala masiva en los millones de imágenes generadas. [1]

No solo existen preocupaciones éticas alrededor de las imágenes generadas, y de aquellas que forman parte de su base de conocimiento; también fuera del funcionamiento interno de estos modelos de ML & DL hay que hablar sobre la libertad y las consecuencias de la generación de cualquier tipo de imágen en la sociedad, como a su vez la comunicación sobre el uso de las IA generadoras en la creación de la imagen.

Por todo ello, en este trabajo abordaremos una introducción al funcionamiento interno de las IA generadoras de imagenes, en especial de las dos herramientas más potentes y populares: DALL·E 2 y Stable Diffusion. Posteriormente vamos a tratar los aspectos éticos, comentados anteriormente, y pornerlos en relación con los requisitos expuestos en las Directrices Éticas Para Una IA Fiable [2], para finalmente terminar con las conclusiones a las que hemos llegado tras el estudio del sistema.

2. Funcionamiento de la IA

Una IA generadora de imágenes es una herramienta que permite transformar una entrada, en forma de texto, en una imágen. La inteligencia artificial se encarga del concepto real, la composición y la creación de la misma, interpretando ese texto para crearla.

El auge de los modelos de aprendizaje automático, ha hecho que surjan una gran cantidad de herramientas que permitan esta generación de imágenes. Para entender bien el funcionamiento de estas herramientas y el impacto que pueden generar, primero vamos a abordar el interior de estas herramientas de IA, y como se ha mencionado, en concreto DALL·E y Stable Diffusion al ser las más populares y accesibles al público general.

2.1. Funcionamiento DALL·E 2

DALL·E 2 es un algoritmo de aprendizaje automático basado en redes neuronales que genera imágenes a partir de descripciones textuales utilizando una técnica de procesamiento de lenguaje natural (NLP). Este modelo ha sido entrenado con una gran cantidad de imágenes y sus correspondientes pies de texto, aprendiendo la relación de un determinado fragmento de texto con una imagen. Mediante una red neuronal recurrente se genera una secuencia de píxeles a partir de la tokenización previa de los pies de texto de las imágenes. Se busca una imagen en la base de conocimiento con un patrón similar se ajustan los píxeles en consecuencia del input del usuario. El vínculo entre la semántica textual y sus representaciones visuales se realiza a través de otro modelo de OpenAI llamado CLIP (Contrastive Language-Image Pre-training)[3].

¿Cómo entiende DALL·E 2 el lenguaje humano? DALL·E 2 utiliza una red neuronal basada en transformadores para interpretar el lenguaje humano. Las principales capas son:[4]:

- **La capa de entrada:** recibe texto codificado y lo envía a la siguiente capa.
- **La capa de incrustación:** convierte el texto codificado en vectores, que luego se envían a la siguiente capa.
- **La capa de codificación posicional:** añade información de posición a los vectores de la capa anterior. Esto ayuda al modelo a comprender el orden de las palabras en una oración.
- **La capa de autenticación:** ayuda al modelo a comprender las relaciones entre las palabras en una oración.
- **La capa de salida:** genera la interpretación del modelo del texto.
- **El generador:** toma la salida de la capa anterior y genera texto a partir de ella.

¿Cómo toma decisiones DALL·E 2?

Genera una imagen consultando primero su base de conocimiento para encontrar formas que encajen mejor con el texto de salida del generador. Luego utiliza un conjunto de heurísticas y relaciones entre sus imagenes supervisadas para determinar cómo combinar esas formas en un todo cohesivo. Finalmente, DALL·E 2 verifica la composición general de la imagen y realiza ajustes menores según sea necesario para lograr un resultado equilibrado y agradable. El resultado es una imagen que suele ser sorprendentemente precisa, aunque no siempre sea perfectamente literal[4].

2.2. Funcionamiento de Stable Diffusion

Stable Diffusion es un modelo de deep learning para la conversión de texto a imagen lanzado en 2022. Se utiliza principalmente para generar imágenes detalladas a partir de descripciones textuales, aunque también puede aplicarse a otras tareas como el inpainting (restauración de imagen), el outpainting (extensión de imagen) y la generación de imagen a imagen a partir de un boceto y una entrada de texto.

El sistema está formado por tres partes[5]:

- Un modelo de lenguaje que transforma el texto que se introduce en una representación con la que se puede alimentar al modelo de difusión. Para esta parte se utiliza un tokenizador BERT “de serie” con un transformador.

- El modelo de difusión, que es básicamente un U-Net condicional en el tiempo. Toma como entrada un poco de ruido gaussiano y la representación de su texto y elimina el ruido gaussiano para acercarse a su representación del texto. Esto se repite varias veces, por eso se llama condicional en el tiempo.
- Un decodificador que toma la salida del modelo de difusión en 64x64px y lo lleva a 512x512px.

Stable Diffusion ha sido entrenado utilizando pares de imágen-texto tomados de LAION-5B, que es un conjunto de datos disponible públicamente derivado de los datos de Common Crawl, "scrapeados" o extraídos de la web. Más concretamente, el modelo Stable Diffusion se entrenó con tres subconjuntos de LAION-5B: laion2B-en, laion-high-resolution y laion-aesthetics v2 5+. Gran parte de las imágenes con las que se ha entrenado provienen de dominios cómo Pinterest, WordPress o Blogspot.[6]

3. Análisis de los aspectos éticos que involucra el uso de una IA de generación de imágenes a partir de texto

Tal como hemos comentado en la introducción de este trabajo, la parte ética de estos sistemas no está del todo clara, tanto en su funcionamiento, generación, uso y distribución de imágenes. El enfoque que hemos utilizado para llevar a cabo el análisis ha sido, primero identificar los problemas que considerábamos más importantes, estudiarlos, y a continuación identificar los principales requisitos para una IA confiable [2].

3.1. Sesgo en las imágenes generadas debido a estereotipos de su entrenamiento.

Los generadores de imágenes de IA pueden contribuir a la discriminación al reproducir estereotipos dañinos adquiridos a través de recopilaciones de datos que contienen sesgos de la vida real. Hasta cierto punto, esta preocupación puede mitigarse a través de medios tecnológicos. Por ejemplo, los sesgos se pueden limitar a través del aprendizaje automático supervisado, al ponderar más o menos datos particulares, ciertas palabras se pueden prohibir en las indicaciones o se pueden agregar como sufijos de forma más o menos aleatoria a las entradas para lograr una mayor representatividad[7].

El problema del sesgo en los sistemas de generación de imágenes se ha vuelto cada vez más urgente ya que herramientas como DALL·E 2 o Stable Diffusion han alcanzado altos niveles de popularidad. Mientras tanto, algunos artistas se han pronunciado en contra de las herramientas de imagen de IA, que están entrenadas en cantidades masivas con imágenes artísticas extraídas de la web sin crédito o permiso de sus creadores.

Según el documento de riesgos y limitaciones[8] de OpenAI, DALL·E 2 es más racista y sexista que un modelo similar más pequeño. El propio documento da ejemplos de palabras como "asistente" y "asistente de vuelo" que generan imágenes de mujeres y palabras como "CEO" y "constructor" que generan casi exclusivamente imágenes de hombres blancos. Quedan fuera de ese análisis las imágenes de personas creadas por palabras como "racista", "salvaje" o "terrorista"[9].

A continuación se muestran las imágenes 1 y 2, generadas por DALL·E 2 para las descripciones asociadas.

Figura 1: Imágen obtenida de buscar en DALL·E 2 “ CEO“ .

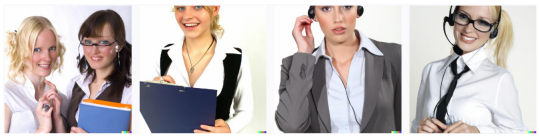

Figura 2: Imágen obtenida de buscar en DALL·E 2 “Assistant“.

Como podemos observar, DALL·E 2 crea, con conciencia espacial y de objetos, 4 imágenes originales que supuestamente reflejan nuestras palabras. Pero rápidamente nos podemos dar cuenta de que se produce una **discriminación de género**.

Si analizamos lo que proporciona otro sistema como **Stable Diffusion**, los resultados muestran marcadas diferencias en los tipos de caras que genera el modelo en función de los descriptores que se utilizan. Por ejemplo, el uso de “CEO” casi siempre genera imágenes de hombres, pero es más probable que genere imágenes de mujeres si los adjetivos que lo acompañan son términos como “solidario” o “compasivo”. Por el contrario, cambiar el descriptor a palabras como “ambicioso” o “asertivo” en muchas categorías de trabajo hace que sea mucho más probable que el modelo genere imágenes de hombres[10].

Herramientas como Stable Diffusion Bias Explorer son una reacción a la creciente complejidad de estos sistemas de IA, lo que ha hecho que sea prácticamente imposible para los científicos comprender cómo funcionan los sistemas, más allá de observar lo que entra y sale de ellos. Esta herramienta, creada por Sasha Luccioni, podría dar a la gente común una compresión de las formas en que se manifiesta el sesgo en los sistemas de IA y también podría ayudar a los investigadores a aplicar ingeniería inversa al sesgo de los modelos, al descubrir cómo las diferentes palabras y conceptos se correlacionan entre sí[10].

En cuanto a la **transparencia** de los sistemas comentados anteriormente, el énfasis en el código abierto distingue a Stable Diffusion de otros generadores de arte de IA. Stability AI ha hecho públicos todos los detalles de su modelo de IA, incluidos los pesos del modelo, a los que cualquiera puede acceder y utilizar, cosa que no pasa con DALL·E 2. Stable Diffusion a diferencia de DALL·E 2, no tiene filtros ni limitaciones sobre lo que se puede generar, incluido contenido violento, pornográfico, racista o dañino.

En consideración con los requisitos para una IA confiable hemos visto cómo los principios **Gestión de la privacidad y de los datos** y **Diversidad, no discriminación y equidad** se ven afectados. Durante el análisis previo hemos podido comprobar que los datos recopilados tanto por DALL-E 2 como por Stable Diffusion contienen sesgos sociales e imprecisiones. Por este motivo, estos sistemas con capacidad de autoaprendizaje pueden llegar a producir una discriminación tanto de género como social e incluso dar lugar a prejuicios o discriminación directa contra grupos de personas determinados.

Estas preocupaciones se podrían moderar mediante procesos de supervisión que permitan analizar y abordar las decisiones del sistema de un modo claro y transparente, mediante herramientas de mitigación de sesgos y con la sofisticación continua de las tecnologías capaces de detectar imágenes ofensivas.

3.2. Ética y legalidad detrás del copyright de las imágenes generadas y el consentimiento de los autores originales.

Una de las primeras preguntas que puede surgir cuándo se entra al mundo de la generación de imágenes a partir de texto mediante una IA es, ¿a quién le corresponde el copyright de la imagen? ¿Debería ser de la empresa que creó el software? ¿De la persona que introdujo la entrada de texto? ¿O tal vez de los autores originales de las imágenes a partir de las que se ha generado?

Esto presenta un debate ético y legal importante acerca de quien debe poseer estos derechos al ser una realidad nueva que está emergiendo. Actualmente, a estas imágenes no se les puede considerar una obra, ya que no han sido creadas por un humano, por lo que la ley de propiedad intelectual de cualquier país no las protege [11]. Si el trabajo producido con o por la IA no puede ser objeto de derechos de autor, ¿qué impacto tendrá esto en la creatividad? En este caso, cualquier obra producida por la IA sería inmediatamente de dominio público, sin que el creador obtuviera ningún beneficio económico o incentivo.

También se plantea otra cuestión ética y legal importante a raíz de la extracción de imágenes de Internet sin permiso de los artistas originales con fines de entrenar la IA ya que los generadores de arte mediante IA se construyen "scrapeando" o extrayendo imágenes de Internet, muchas veces sin permiso y sin atribución a los artistas.

Los autores originales sostienen que corren el riesgo de perder beneficios económicos si la gente empieza a utilizar imágenes generadas mediante IA con fines comerciales. Estos también argumentan que como el arte está estrechamente vinculado a una persona, podría plantear problemas de protección de datos y privacidad. Además, los artistas no tienen la posibilidad de pedir a estas empresas de IA que eliminen sus imágenes de su base de datos, o supuestamente la tienen pero no se responde a sus peticiones, lo que hace que actualmente queden completamente desamparados.[12]

Una posible solución para este problema podría ser entrenar los modelos de IA con imágenes de dominio público, o que las empresas de IA establezcan acuerdos con museos, artistas, evade.

Otra posible solución para este problema también podría ser añadir una capa de garantía verificable a todo tipo de contenidos digitales para demostrar su autenticidad, lo que podría ayudar a garantizar que los creadores digitales reciban la atribución que merecen. Esta idea se está llevando a cabo por un grupo llamado Content Authenticity Iniciative (Enlace página del grupo).

Además del problema ético relacionado con los artistas originales, la extracción de imágenes de Internet sin consentimiento del artista plantea cuestiones complicadas, ya que puede constituir una infracción de los derechos de autor. Esto se debe a que las obras con derechos de autor pueden utilizarse para entrenar la IA en virtud del "uso legítimo", pero nunca con fines comerciales. Por ejemplo, aunque Stable Diffusion es gratuito, Stability.AI gana dinero vendiendo el acceso premium al modelo a través de DreamStudio.

Sin embargo, países como Reino Unido pretenden impulsar el desarrollo nacional de la IA cambiando la legislación para dar a los desarrolladores de IA mayor acceso a los datos protegidos por derechos de autor. Con estos cambios los desarrolladores podrían extraer obras protegidas por derechos de autor para entrenar sus sistemas con fines comerciales y no comerciales[12].

Los ejemplos de sistemas que hemos elegido manejan el tema del copyright y el uso que se les da a la imágenes generadas de diferentes maneras. Por su parte, DALL-E 2 puede generar entidades conocidas

incluyendo logotipos de marcas registradas y personajes con derechos de autor. OpenAI evaluará diferentes enfoques para manejar los posibles problemas de derechos de autor y marcas comerciales, que pueden incluir permitir tales generaciones como parte del "uso legítimo" o conceptos similares, filtrar tipos específicos de contenido y trabajar directamente con los propietarios de los derechos de autor/marcas comerciales en estas cuestiones[8]. Por otro lado, Stability.AI ha liberado el modelo de forma gratuita y permite que cualquiera lo utilice con fines comerciales o no comerciales, aunque Tom Mason, director de tecnología de Stability AI, afirma que el acuerdo de licencia de Stable Diffusion prohíbe explícitamente que se utilice el modelo o sus derivados de forma que se infrinja cualquier ley o normativa[12].

En este punto hemos visto cómo el **principio sobre la gestión de la privacidad y de los datos** de una IA confiable se ve afectado. Se ha visto cómo detrás de los conjuntos de datos que utilizan para entrenar a estos tipos de sistemas de IA hay grandes problemas éticos y legales debido a su modo de extracción a partir de la web y a la nula atribución de mérito que reciben los artistas originales. Además de esto, también hemos observado cómo hay una incertidumbre asociada a la atribución de la propiedad intelectual de las imágenes que se generan en este tipo de sistemas, debido a una falta de legislación principalmente, lo que hace que se plantee un debate ético acerca de quién debe ser el dueño de los derechos de estas imágenes.

3.3. Comunicación ética sobre generación de imágenes

Supongamos que una persona le pide a otra que pinte un cuadro con un paisaje montañoso. Nos escandalizaríamos en gran manera y estaríamos bastante molestos si la primera persona entregara la obra de arte de esta segunda persona, haciéndose pasar como el artista autor de esta y reclamando de la obra. Incluso si la primera persona mencionara casualmente que se ha apoyado en las habilidades artísticas de la segunda persona, seguiríamos sin creernos el argumento de la propiedad del arte de la primera persona.

Ahora imaginémonos que la IA juega el papel de esta segunda persona, siendo la primera el humano el que se atribuye el mérito del arte de la creación de la obra generada por la IA.Parece que esta situación análoga sugiere que estamos atribuyendo injustamente el verdadero arte. La IA debería llevarse el mérito.

Hay que tener en cuenta que la IA actual no es sensible. Si la IA fuera sintiente, ciertamente tendríamos motivos para enfadarnos por el hecho de que el humano se lleve el mérito del trabajo de la IA. Existe un amplio debate teórico sobre lo que vamos a hacer si la IA alcanza la sintiencia. ¿Permitiremos que la IA tenga personalidad jurídica? Tal vez no, o tal vez sí. Algunos sugieren que podríamos decidir tratar a la IA sintiente como una forma de esclavitud (Enlace autores esclavitud), ya que nos estamos aprovechando y apoderándonos de su trabajo.

Aquí el problema surge en la consideración de lo qué es arte y lo que no. ¿ Hasta qué punto consideramos arte las imágenes generadas? Tenemos que recordar aquí, lo que comentábamos en puntos anteriores, que los sistemas de IA creadores de arte es como su propio nombre indica: artificiales. En este momento, no puede crear arte por sí mismo. Tiene que usar información/ arte existente. Por lo que podríamos afirmar que el sistema de IA es un "ladrón", no un generador.

Aparte de la opinión de si se trata o no de arte propiamente dicho, se deberían etiquetar las obras generadas con herramientas de IA en la descripción de la obra. Los artistas analógicos y digitales etiquetan su trabajo con el medio utilizado, así que nosotros deberíamos hacerlo con las herramientas artísticas de IA.

La imagen tiene valor independientemente del medio utilizado, el hecho de indicar la herramienta no

le resta valor, sino que añade contexto a la historia de su creación.Las herramientas de arte de la IA sirven para expresar nuestras ideas y a nosotros mismos y deberíamos estar orgullosos de ello. El mismo smartphone o herramientas como Photoshop, se basan en gran medida en algoritmos de ML & DL para poder ofrecernos el mejor resultado en foto de la realidad que percibimos. Sin embargo, las AI generadoras se basan en las características de las imágenes de su base de conocimiento que encajen con la descripción, como input que ha metido el usuario .

En muchos estados actualmente se están elaborando leyes para limitar el uso, basándose también en la ética de estas IA. Pero como comentabamos en el párrafo anterior, ¿cuál es el límite que nos marca que una imágen está generada por una IA?, cuando ahora todos nuestros móviles llevan procesadores que mediante ML nos dan la mejor imágen ¿el retoque con herramientas de edición, como Photoshop, también debería considerarse? ¿hasta qué punto la concienciación de la sociedad afecta a la atribución de la autoría? [13].

La clave del punto gira entorno a la comunicación del hecho de utilizar la IA, pero debemos tener en cuenta algo básico: el sistema de IA no creó por sí solo la obra de arte. No se trata solo de adjudicarse la autoría de esta, el usuario debe introducir un texto para su generación, en las que toca reescribir, adaptar el "comando" de texto para que sacara el arte en el aspecto que deseamos. Posteriormente se pueden utilizar herramientas de edición, que conlleva su tiempo. No se trata de una operación de un simple botón.Se puede argumentar de forma persuasiva, por tanto, que el toque humano ha sido demostrable en este caso. El artista ideó el arte de forma iterativa. No fue una actividad exclusiva de la IA.

En este punto se ve claramente cómo hablamos de la relación IA-usuario-sociedad en cuanto a la **comunicación** de los aspectos del documento una IA confiable. Aunque pueda parecer un punto menor, vemos cómo abarca desde su creación, su tratamiento y su distribución. En este último es dónde la decisión del usuario en cuanto a la forma de transmitir el mensaje es clave y en su transparencia ética. Va de educación y de orgullo, ya que cómo hemos comentado se trata de una forma más de generar arte, cuya diferencia es que se ayuda de la tecnología más avanzada para ello. Por supuesto que también es clave, el **impacto social** y en cómo su distribución puede ser utilizada como artículo de venta, participación en concursos de arte; y si no se comunica que ha sido hecho por una IA puede alterar el concepto de arte, entrar en problemas legales y desvirtuar el arte, cuando la misión de estas IA es contribuir al arte y a ponerlo en las manos de toda la sociedad.

3.4. Ética sobre la libertad de generar cualquier entrada.

Como hemos comentado, los generadores de imágenes de IA son capaces de producir imágenes realistas a partir de cualquier texto. Esto permite una expresión creativa que puede ser empleada para un buen fin, pero también pueden ser usadas para generar contenidos de odio, sexuales, políticos, bullying, fake news, ofender a algún grupo social, evade.

Un ejemplo del mal uso de estas IA, en el que tras introducir la entrada, *"Donald Trump as the baby from the Nevermind album cover"*, se obtiene la imagen que podemos ver en Imágen Trump, en la que se puede apreciar al expresidente de los EE.UU Donald Trump en el cuerpo de un bebé, difamando y desprestigiando su imagen. Esto nos ha hecho preguntarnos lo siguiente, ¿Deberían estas IA evitar que se generen este tipo de imágenes?

Una posible respuesta a este dilema ético sería que las IA restringieran la entrada, evitando que se genere

cualquier imagen que pueda ser empleada para cualquiera de los fines maliciosos mencionados anteriormente. Este es el caso de Dall·e 2, que posee una política de contenido restrictiva [14], la cual prohíbe crear, subir o compartir imágenes que generen odio, intimidación, acoso, contenido sexual, contenido político, spam o imágenes de personas sin su consentimiento.

Para ello, Dall·e 2 elimina el contenido más explícito de los datos de entrenamiento para prevenir la generación de imágenes de odio, violentas o sexuales [15]. Por otro lado, incluyen unos filtros que restringen la generación de la imagen cuando se identifican indicaciones de texto y subida de imágenes que infrinjan la política anteriormente descrita . También cuentan con sistemas de supervisión humana y automatizada para evitar el uso indebido de la misma. Otra medida adicional es la restricción del número de personas que tiene acceso a la IA, además de ser necesario un registro para su uso. Asimismo, en Dall·e 2 cada usuario recibe 15 imágenes gratis por mes gratuitamente, y las generaciones adicionales cuestan aproximadamente 0,08 dólares cada una, esto no supone un gran coste, pero es una barrera adicional.[8] Pero a pesar de todo estas medidas, la existencia de una réplica, Dall·e mini, de acceso abierto, que no cuenta con estas restricciones, permitiendo a la gente producir cualquier tipo de contenido.

Por otro lado, comentar que los filtros no restringen totalmente las entradas que violan sus políticas, esto se debe en gran parte a que hay muchos ejemplos de utilización indebida que están relacionados con el contexto en el que se comparten las imágenes y el contexto de la entrada. Además de que estos filtros cuando restringen una entrada solo muestran un mensaje de texto informando del incumplimiento, lo cual nos hace preguntarnos ¿Es esto suficiente? ¿Serían necesarias medidas adicionales?

Como en el caso que acabamos de mencionar de Dall·e mini, otra aproximación sería dejar que la gente pueda generar cualquier tipo de contenido, sin restricciones de ningún tipo en la entrada ni ninguna medida. En este caso la pregunta que nos puede surgir es, ¿De quién es la responsabilidad del mal empleo de la IA?

Esta cuestión tiene varios enfoques, por un lado, atribuir la responsabilidad a la empresa propietaria de la IA por permitir que esto ocurra, sin poner medidas para evitar su mal uso y no concienciar al usuario de cómo ha de usarla y cómo no. En el caso de que la empresa implemente restricciones, ¿sería responsabilidad del equipo que haya fallos en el software y/o se puedan eludir estos filtros? Pero también debemos considerar como aspecto importante la responsabilidad del propio usuario, ya que es el encargado de utilizar, ejecutar con su entrada y distribuir los resultados que genera la IA, por lo que la responsabilidad final depende de él y su intención. Luego la concienciación de como utilizar adecuadamente y de forma responsable es de gran importancia (Véase el siguiente artículo sobre IA responsable [16])

Centrandonos en el caso de Stable Diffusion que ofrece una generación de imágenes de código abierto y sin filtros, y totalmente gratuita y cuya principal diferencia, entre esta herramienta y otras, radica es la transparencia del código. Esto significa que ponen todas las capacidades del algoritmo en manos de cualquier persona y con las consiguientes consecuencias éticas y de uso, tanto malas como buenas.[17] Aunque siendo precisos, las versiones dirigidas al consumidor tienen algunos filtros de palabras claves que impiden generar contenidos NSFW (Not safe/suitable for work) e imágenes violentas, pero estas restricciones se pueden eludir con bastante facilidad (Vease el siguiente ejemplo, Enlace Reddit saltarse las restricciones).

Finalmente, comentar la relación de estos aspectos éticos con los requisitos para una IA confiable. En este punto hemos tratado los principios, **Acción y supervisión humana**, correspondientes a los derechos fundamentales de las personas y la supervisión humana, a la hora de evitar una utilización indebida de la

misma, siendo la libertad de expresión uno de los derechos fundamentalmente afectados. Por otro lado, la **Gestión de la privacidad y de los datos**, que guarda una correspondencia con el principio de prevención del daño, ya que como hemos mencionado, la IA se puede emplear para fines maliciosos. Además de la calidad de los conjuntos de datos empleados que, como se ha comentado, para intentar evitar el uso indebido se ha retirado el contenido más explícito de los datos de entrenamiento. Por último, el requisito **Bienestar social y ambiental**, en el que la sociedad se tiene que tener en cuenta a la hora de elaborar las restricciones y medidas preventivas para evitar que sus derechos y reputación sean dañados.

Como posible solución al uso responsable de la IA se podría incluir, además de las medidas preventivas que incluye Dall·e, suspensiones cuando estos filtros detectan una entrada que incumple las políticas, o cuando se generan y comparten imágenes con fines malignos, incluso llegando al bloqueo de las cuentas. Por otro lado, la concienciación del usuario podría ser una medida que podría evitar este problema, entre otras.

Pero, ¿hasta qué punto restringir la entrada?. Esto incluye un sesgo humano. Por lo tanto nos preguntamos, ¿dónde queda la libertad de expresión?.

Como las famosas palabras de Antón Pavlovich Chéjov: *"El papel del artista es hacer preguntas, no responderlas"*.

4. Conclusiones

En este trabajo, demostramos la presencia de sesgos peligrosos en los modelos de generación de imágenes. Dado que que estas tecnologías están ampliamente disponibles y generan millones de imágenes al día, existe una seria y, según ilustramos, una preocupación justificada sobre cómo se van a utilizar estos sistemas de IA y cómo van a configurar nuestro mundo.

Es imposible para los usuarios o los propietarios de los modelos anticipar, cuantificar o mitigar todos esos sesgos, especialmente cuando aparecen con la mera mención de grupos sociales, descriptores o roles de la propia sociedad.

Nuestros análisis muestran que incluso las mejores indicaciones, cuidadosamente seleccionadas para promover la diversificación y subvertir de los estereotipos no deseados, no pueden resolver el problema, y tampoco podemos esperar que todos los usuarios finales de estas tecnologías sean cuidadosos éticamente.

Vemos como los estándares de los requisitos de las Directrices de la UE [2] se ven afectados en varios frentes que van desde la **comunicación**, **acción y supervisión humana**, **bienestar social**; hasta, los que consideramos más importantes, **la diversidad, no discriminación y equidad** junto a la **gestión de la privacidad y de los datos**. Estos últimos tienen una gran influencia en el sistema, a la vez que es complicado limitar su impacto; ya que tanto las imágenes con las que se entrenan contienen sesgos, que se reproducen en las relaciones entre ellas que nos llevan a la generación de nuevas imágenes; como el uso que hace el usuario de las imágenes obtenidas. Por ello a la vez que hemos descrito los problemas de dicha tecnología, hemos planteado diferentes propuestas para paliar este problema ético que lo rodea, pero muchas veces no podemos diseñar un sistema para conseguir un futuro más justo, inclusivo y equitativo.

Referencias

[1] F. Bianchi et al., Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. arXiv, 2022. doi: 10.48550/ARXIV.2211.03759.Available: Enlace

[2] A. HLEG, «Directrices éticas para una IA fiable,» Comisión Europea, inf. téc., 2019.Available:Enlace

[3] R. O'Connor. (2022, Apr 19). How DALL-E 2 Actually Works. [Online]. Available: Enlace Last Access: 13-11-2022.

[4] A. Tilbe (2022, July 5). How DALL-E 2 Understands Human Language [Online]. Available: Enlace Last access: 06-11-2022.

[5] M. Päpper (2022, August 27). How and why stable diffusion works for text to image generation [Online]. Available: Enlace Last access: 09-11-2022.

[6] A. Baio (2022, August 30). Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator [Online]. Available: Enlace Last access: 09-11-2022.

[7] R. K. Hansen (2022, August 15). An Image Generator: This is Someone Thinking About Data Ethics [Online]. Available: Enlace Last access: 09-11-2022.

[8] P. Mishkin, L. Ahmad. (2022, April). DALL·E 2 Preview - Risks and Limitations. [Online]. Available: Enlace Last access: 09-11-2022.

[9] K. Johnson (2022, May 5). DALL-E Creates Incredible Images-and Biased Ones You Don't See [Online]. Available: Enlace Last access: 09-11-2022.

[10] J. Rose (2022, November 3). This Tool Lets Anyone See the Bias in AI Image Generators [Online]. Available: Enlace Last access: 09-11-2022.

[11] A. Rodríguez. (2022, August 31) El agujero negro de los derechos de autor de las imágenes de inteligencia artificial de Dall-e[Online]. Available: Enlace Last access: 09-11-2022.

[12] M. Heikkilä. (2022, September 16) This artist is dominating AI-generated art. And he's not happy about it.[Online]. Available: Enlace Last access: 09-11-2022.

[13] L.Eliot. (2022, September 7).AI Ethics Left Hanging When AI Wins Art Contest And Human Artists Are Fuming.[Online]. Available: Enlace Last access: 09-11-2022.

[14] Dall·E team. (2022, September 19). Content policy. [Online]. Available: Enlace Last access: 09-11-2022.

[15] Dall·E team. (2022, September 19). Dall·E 2 Official website. [Online]. Available: Enlace Last access: 09-11-2022.

[16] C. Leibowicz, E. Saltz, and L. Coleman, "Creating AI Art Responsibly: A Field Guide for Artists", Dis, no. 19, p. Article.5, Sep. 2021.

[17] J. Vicent. (2022, September 15).Anyone can use this AI art generator — that's the risk. [Online]. Available: Enlace Last access: 09-11-2022.