Bonn, Bucarest, Dublín, Lisboa, Madrid, Milán, París, La Haya, Viena, Varsovia

IA Generativa: Implicaciones para la protección de datos

Grupo de trabajo sobre IA de CEDPO

16 de octubre de 2023

Guía creada por CEDPO originalmente en inglés.
Traducción al castellano mediante herramientas de traducción automatizada con revisión pormenorizada por parte de APEP.
En caso de duda consultar original en https://cedpo.eu/publications/

Acerca de esta Guía

La inteligencia artificial no es un concepto nuevo para los DPO y los profesionales de la protección de datos. Sin embargo, la IA generativa sí lo es. Cuando se lanzó ChatGPT de OpenAI en noviembre de 2022, la mayoría de los profesionales de la protección de datos nunca habían oído hablar de la IA generativa y, desde luego, no se ocupaban de este tipo de tecnologías en su trabajo diario.

Ahora, con ChatGPT en manos de más de 100 millones de usuarios en todo el mundo, y muchos otros proveedores como Google Bard y Claude de Anthropic entrando en el mercado, se ha convertido en una realidad operativa, y una necesidad, para los profesionales de la protección de datos hacer frente a las consecuencias de las herramientas de IA generativa que se están utilizando rápidamente en las organizaciones. Tanto si estas herramientas se adoptan de forma simple como si las organizaciones las perfeccionan utilizando sus propios conjuntos de datos, existen implicaciones novedosas y aún no examinadas para la protección de datos, que los profesionales de la protección de datos deben asumir rápidamente.

El objetivo de este documento es guiar a los profesionales de la protección de datos a través del laberinto de cuestiones que están surgiendo a medida que estas tecnologías se adoptan rápidamente en las organizaciones. Entre otras cuestiones clave, este documento examina los riesgos del intercambio de datos, la exactitud de los datos personales, la realización de EIPD sobre herramientas de IA generativa, la aplicación de la protección de datos desde el diseño, la selección de una base legal para la formación de sistemas de IA generativa, la optimización de las estructuras organizativas, la aplicación de técnicas de mejora de la privacidad y la gestión de los derechos de los interesados en el contexto de estas tecnologías.

No habrá futuro sin IA generativa, y dado que los datos desempeñan un papel tan fundamental en la formación y el funcionamiento de estos sistemas, los responsables de la protección de datos desempeñarán un papel central a la hora de garantizar que tanto las normas de protección de datos como las de gobernanza de datos estén en el centro de estas tecnologías.

CEDPO
CONFEDERATION OF EUROPEAN DATA PROTECTION ORGANISATIONS

Índice

1. Exactitud de los datos personales

La exactitud de los datos personales tratados por herramientas de inteligencia artificial generativa (IA) es una cuestión fundamental de la protección de datos con este tipo de tecnologías. El artículo 5, apartado 1, letra d), del RGPD establece que "los datos personales serán exactos y, en caso necesario, se mantendrán actualizados; se adoptarán todas las medidas razonables para que los datos personales que sean inexactos (...) se supriman o rectifiquen sin dilación".

Es obvio, y una cuestión de sentido común, que el tratamiento de datos personales inexactos puede tener implicaciones muy reales para el titular de los datos que hay detrás de ellos. Sin embargo, las herramientas de IA generativa, como el chatbot basado en texto de OpenAI, ChatGPT, ampliamente utilizado, tienen intrínsecamente numerosas inexactitudes en los datos que procesan. Por su propia naturaleza, estas herramientas ingieren grandes cantidades de datos de entrenamiento procedentes de ejercicios masivos de extracción de datos en Internet. Necesariamente, estos datos vienen con todas sus imperfecciones, y se convierten en parte del banco de datos que los usuarios de ChatGPT consultan. Cuando un usuario recibe una respuesta que es total o parcialmente inexacta, se genera lo que los proveedores de IA llaman "alucinaciones" o, en la jerga, "falsedades".

Incluso la propia OpenAI, en su página web, advierte a los usuarios de los peligros que entraña y de que no se puede confiar automáticamente en la exactitud de los datos recuperados. En una sección titulada "Limitaciones" señala que "ChatGPT a veces escribe respuestas plausibles pero incorrectas o sin sentido"[1] . Para agravar el problema, OpenAI señala además que la herramienta suele aumentar las imprecisiones al adivinar esencialmente lo que quiere decir un usuario inseguro. Afirma: Lo ideal sería que el modelo hiciera preguntas aclaratorias cuando el usuario hiciera una consulta ambigua. En cambio, nuestros modelos actuales suelen adivinar la intención del usuario.[2] '

Si a esto le añadimos que las condiciones de tratamiento de datos de ChatGPT dejan claro que el usuario es el responsable del tratamiento de los datos, mientras que OpenAI es simplemente el procesador de datos, debería quedar claro, para los usuarios, que este es un mercado en el que "el comprador debe tener cuidado". ¿Por qué? Porque si alguna de las partes sigue tratando datos personales inexactos, se convertirá en responsable de cualquier incumplimiento del artículo 5 (1) (d) anterior. En el contexto de ChatGPT, por tanto, basarse en datos personales inexactos proporcionados por la herramienta hará que el usuario sea responsable del incumplimiento del RGPD, especialmente cuando dicha reutilización afecte a los derechos y libertades fundamentales de los interesados.

[1] https://openai.com/blog/chatgpt
[2] Ibid.

Las organizaciones deben entender que esto no es meramente un punto teórico y que los reguladores ya han pedido cuentas a las empresas de IA generativa sobre la exactitud de sus datos. En marzo de 2023, la Autoridad Italiana de Protección de Datos bloqueó el despliegue de ChatGPT en Italia, señalando, entre otras cosas, que los datos a menudo no eran exactos. Señaló, basándose en "las pruebas realizadas hasta la fecha, que la información facilitada por ChatGPT no siempre coincide con las circunstancias reales, de modo que se procesan datos personales inexactos".[3]

Así pues, los responsables de la protección de datos (DPO deben ser conscientes de los riesgos que entraña el tratamiento de datos inexactos. Los usuarios de la organización de un DPO deben recibir directrices claras que les ayuden a comprender que los resultados de cualquier herramienta de IA generativa, como ChatGPT, van acompañados de una advertencia sanitaria, a saber, que el usuario humano sigue siendo el responsable último de verificar la exactitud de cualquier dato personal obtenido. Este es un punto crítico.

Otro riesgo relacionado procede de la segunda cláusula de la letra d del apartado 1 del artículo 5, según la cual los datos personales "se mantendrán actualizados". ChatGPT, y herramientas similares como Bard de Google y Claude de Anthropic, se basan en actividades de "scraping" hasta un determinado momento, lo que significa que su banco de datos se desactualiza y, por tanto, con el tiempo, no responden necesariamente a acontecimientos actualizados. Esto crea el riesgo evidente de que los usuarios obtengan datos personales que ya no son pertinentes, o tal vez carecen de contexto, o son simplemente inexactos, dado cómo han cambiado los acontecimientos o cómo ha avanzado la información en el período intermedio.

Los responsables de la protección de datos también deben ser conscientes de las formas en que la parcialidad y la discriminación no mitigadas en los conjuntos de formación podrían conducir indirectamente a resultados de datos inexactos, abriendo de nuevo al usuario a los riesgos de seguir procesando datos inexactos.

Un último riesgo global de los chatbots de IA generativa es el tono que adoptan: un nivel oracular de certeza y autoridad que casi podría calificarse de patrón oscuro, tan engañoso es en su efecto sobre la evaluación de los resultados de búsqueda. Cuando los chatbots de IA generativa se equivocan o son inexactos de forma palpable, a menudo lo hacen de una manera muy segura y confusamente definitiva, una actitud que enmascara el hecho de que, como admite OpenAI, por ejemplo, la respuesta puede ser simplemente un "sinsentido". En cualquier resultado de búsqueda, el tono de la respuesta debe ignorarse y, de nuevo, los usuarios deben darse cuenta de que el resultado de estas herramientas requiere una evaluación humana, sobre todo cuando se trata de cuestiones sobre la exactitud de cualquier dato personal implicado.

[3] h ttps://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9870847#english

2. Compartir datos personales con herramientas de IA generativa

La Inteligencia Artificial (IA) ha pasado rápidamente de ser un concepto de ciencia ficción a una característica relativamente común de nuestra vida. Una rama de la IA que está emergiendo con rapidez es la IA generativa, que puede crear datos nuevos, previamente inexistentes, que se asemejan mucho a los datos de entrada. En las condiciones adecuadas, los modelos de IA generativa pueden generar textos, imágenes, música, etc., de gran calidad. Sin embargo, la comodidad y el potencial innovador de la IA generativa tienen un coste. A pesar de sus prometedoras capacidades, el intercambio de datos personales con estos sistemas presenta riesgos sustanciales para la privacidad, la confidencialidad y la integridad y seguridad de los datos. Comprender estos riesgos es esencial para proteger los derechos individuales de protección de datos y mantener un entorno digital seguro.

Como la mayoría de los sistemas de IA, la IA generativa se basa en datos. El entrenamiento tradicional de la IA consiste en alimentar grandes conjuntos de datos a modelos de IA que pueden aprender patrones y características a partir de estos datos. Una vez completado el entrenamiento, el sistema de IA está equipado para generar resultados basados en los patrones y características aprendidos. Esto significa que una vez que los datos personales forman parte del conjunto de entrenamiento de la IA, contribuyen a la formación del modelo interno de la IA e influirán invariablemente en su comportamiento y sus resultados. De hecho, los datos se convierten en "parte" de la IA, en el sentido de que informan la comprensión y el conocimiento del sistema. Esto plantea importantes problemas de protección de datos cuando los datos personales se utilizan como datos de entrenamiento.

Los modelos generativos de IA entrenados a partir de datos personales pueden extraer información sensible como nombres, direcciones, información sanitaria o incluso datos financieros, y luego volver a publicar esos datos en los resultados de búsqueda para diferentes usuarios. Además, los modelos generativos de IA pueden amplificar la exposición generando más datos similares a los originales. Terceros pueden entonces explotar estos datos para actividades ilegales, como publicidad invasiva, estafas de phishing o, en casos más graves, fraude o robo de identidad. Esto pone de manifiesto la complejidad de controlar cómo utilizan los datos personales los modelos generativos de IA. Una vez que los datos personales se han compartido con los modelos generativos de IA, la gestión y el seguimiento de su uso se convierten en una tarea intrincada (si no imposible), debido a la naturaleza de cómo los sistemas de IA

procesan la información, así como almacenan y replican los datos en diferentes sistemas. Por lo tanto, retractarse de los datos personales compartidos con modelos de IA generativa puede ser increíblemente difícil o poco realista. La lección para los DPO es que los usuarios *deben* entender con precisión qué tipo de información *puede* y *no puede* compartirse con las herramientas de IA generativa, porque una vez que se comparten los datos personales, se ha cruzado el Rubicón, y será muy difícil deshacer lo que se ha hecho.

Uno de los riesgos más alarmantes asociados al intercambio de datos personales con la IA generativa es la creación y proliferación de "deepfakes". Los deepfakes se refieren a la aplicación de la IA para crear, alterar o manipular contenidos, como imágenes, audio y vídeo, de tal manera que se fabriquen contenidos hiperrealistas pero totalmente falsos. Mediante el entrenamiento con datos personales, la IA generativa puede generar medios sintéticos que suplantan de forma convincente a personas físicas o jurídicas. Estos deepfakes pueden utilizarse con fines maliciosos, como en campañas de desinformación, fraude o acoso. En relación con esto está el hecho de que la precisión de las decisiones de la IA generativa depende en gran medida de la calidad y diversidad de los datos de entrenamiento de entrada. Si estos datos personales están sesgados, los resultados de la IA también pueden estarlo, lo que tendría consecuencias injustas.

La IA generativa es muy prometedora para numerosas aplicaciones, pero su uso de datos personales debe gestionarse cuidadosamente para mitigar los posibles riesgos. Mediante el empleo de controles estrictos de protección de datos, prácticas éticas de IA y protecciones jurídicas sólidas, puede ser posible aprovechar el potencial de la IA generativa, salvaguardando al mismo tiempo los derechos individuales de protección de datos y fomentando un entorno digital seguro.

3. ¿Qué es una base jurídica adecuada?

La base jurídica que se aplica correctamente a la formación de sistemas de IA con datos personales es una consideración clave. A primera vista, no existe un candidato obvio que legitime claramente esta actividad de tratamiento y que al mismo tiempo defienda los derechos de protección de datos de las personas afectadas. Se trata de una consideración crítica porque el volumen de datos de entrenamiento que se utilizan para las aplicaciones de IA generativa es enorme, y su tamaño no hace más que crecer. Para que estas actividades de formación continúen, y para que la IA cumpla sus promesas, no puede basarse en un fundamento jurídico incierto en lo que respecta a los datos personales. Por otra parte, la Ley de Inteligencia Artificial (Ley de IA) no es particularmente instructiva en este punto, dado que el artículo 10 (que trata de la gobernanza de los datos y la gobernanza de los datos de formación para los sistemas de IA) no crea una base jurídica específica para el uso de datos personales para la formación de sistemas de IA. Así pues, debemos acudir al RGPD en busca de una base jurídica adecuada para esta actividad.

En primer lugar, veremos brevemente cómo se utilizan los datos para entrenar sistemas de IA generativa. Esto tiene lugar de cuatro formas generales:

1. Basado en datos personales extraídos de Internet;
2. Cuando los datos personales han sido proporcionados por los usuarios del sistema de IA, como cuando envían indicaciones a las herramientas de IA Generativa;
3. Cuando los datos personales se hayan recogido de terceros, como intermediarios de datos, o de empresas que dispongan de bases de datos pertinentes para la fase de entrenamiento de la inteligencia artificial (por ejemplo, una base de datos de resoluciones judiciales para una herramienta de inteligencia artificial predictiva en el ámbito jurídico); y
4. Cuando los desarrolladores/operadores de IA utilizan los datos personales contenidos en sus propias bases de datos para entrenar el sistema de IA.

En estos casos, de conformidad con el artículo 6 del RGPD, son tres las bases jurídicas más pertinentes: el contrato, el interés legítimo y el consentimiento.

1. **Contrato:**

El artículo 6, apartado 1, letra b, del RGPD señala que el contrato puede constituir una base jurídica para el tratamiento de datos personales cuando dicho "tratamiento sea necesario para la ejecución de un contrato en el que el interesado sea parte o para tomar medidas a petición del interesado antes de celebrar un contrato".

La aplicación de la primera rama de la base jurídica contractual (es *decir,* la ejecución del propio contrato exigiría demostrar que el entrenamiento del sistema de IA (y no el uso de la IA una vez entrenada es estrictamente necesario para la ejecución de un contrato con el interesado.

Este requisito de necesidad es interpretado de forma muy restrictiva por las autoridades de protección de datos. Según el Comité Europeo de Protección de Datos (CEPD), no debería ser posible ejecutar el objeto principal del contrato específico con el interesado si no se produce el tratamiento de los datos personales en cuestión. En otras palabras, el tratamiento de los datos personales debe ser una condición necesaria para la ejecución del contrato.

Teniendo en cuenta esta interpretación restrictiva, hay muy poco margen para la base contractual a la hora de entrenar un sistema de IA. En teoría, esta base podría aplicarse cuando el uso del sistema de IA sea objeto del contrato celebrado entre el operador de IA y el usuario, y cuando no haya otra forma de ejecutar este contrato que entrenar la IA con los datos de los usuarios.

En cuanto a la segunda rama de esta base jurídica, es decir, los pasos precontractuales, su aplicación exigiría demostrar que un interesado hizo una petición en el contexto de la posible celebración de un contrato y que no hay otra forma de satisfacer sus demandas que entrenar (y no sólo utilizar una vez entrenada la IA. Se trata de una opción aún más restrictiva y limitada que la primera parte de esta base jurídica.

En conjunto, las circunstancias en las que la base jurídica del contrato podría utilizarse para justificar la formación de sistemas de IA con datos personales son muy limitadas y, en la práctica, esta base no será una opción viable para fundamentar tales actividades de tratamiento.

En el caso de la IA generativa, el contrato como base jurídica es, en cualquier caso, especialmente inadecuado, dado que normalmente no existe ningún contrato entre los interesados cuyos datos se utilizan y las organizaciones responsables de formar dichos sistemas con esos datos.

2. Intereses legítimos

La base del interé legítimo sólo podría aplicarse a condición de que el responsable del tratamiento de datos complete una evaluación de los intereses legítimos para garantizar que no prevalecen sobre los intereses o los derechos y libertades fundamentales del interesado.

Sin embargo, esto puede suponer un reto, especialmente porque, la mayoría de las veces, la organización que está detrás de la formación de herramientas de IA generativa, como OpenAI, no está en contacto directo con los interesados ni tiene ningún tipo de relación con ellos. En este sentido, cabe destacar las recientes acciones de la Autoridad de Supervisión de Protección de Datos italiana (Garante per la protezione dei dati) contra ChatGPT. En marzo de 2023, la autoridad bloqueó ChatGPT en el territorio italiano hasta que OpenAI pudiera responder satisfactoriamente a ciertas preguntas, una de las cuales era que OpenAI debía especificar la base legal para entrenar a ChatGPT con datos personales. En su respuesta a este punto, OpenAI identificó los intereses legítimos como base jurídica. Se trata de un compromiso y una declaración muy significativos por parte de OpenAI, ya que vincula efectivamente la enorme tarea de entrenar sistemas de IA generativa a una base jurídica que es inherentemente incierta, dado el derecho explícito de los interesados en virtud del artículo 21 del RGPD a oponerse a dicho tratamiento.

Para poder invocar efectivamente la base del interés legítimo sería necesario, en particular:

- un estudio caso por caso del contexto de formación y de utilización de la IA, así como de la recogida de los datos personales utilizados para verificar que el tratamiento de datos responderá a las expectativas razonables de los interesados;
- una demostración de la estricta necesidad de este tratamiento y del hecho de que la IA no puede trabajar eficazmente sin haber sido entrenada con los datos personales en cuestión;
- una mayor transparencia del tratamiento de datos con respecto a los interesados. Toda la información exigida por el RGPD deberá facilitarse a los interesados de forma adecuada;

- un sistema eficaz de exclusión voluntaria puesto en conocimiento de los interesados en un plazo razonable antes de que sus datos se faciliten al sistema de IA[4] ;
- de manera más general, un sistema eficaz para garantizar el respeto de los derechos de los interesados
  que sería difícil de aplicar dadas las particularidades del funcionamiento de la IA generativa.

3. Consentimiento:

La base del consentimiento también podría aplicarse, pero sólo en circunstancias muy claramente delimitadas. Aunque en casos extremos puede ser la única base jurídica posible (por ejemplo, cuando se tratan categorías especiales de datos o datos relativos a menores), por regla general tiene muy poca cabida en el entrenamiento de sistemas generativos de IA, tal como se conciben actualmente. Todo el aparato utilizado para el entrenamiento de los sistemas de IA hace casi imposible obtener el consentimiento. Esto se debe, en primer lugar, a que la mayoría de los datos utilizados para entrenar dichos sistemas se compran a intermediarios de datos que han obtenido estos datos raspando Internet, una actividad que necesariamente no implica la obtención del consentimiento de los interesados subyacentes. De hecho, la propia legalidad del "scraping" de datos como actividad comercial dista mucho de ser segura y la reciente comunicación conjunta de doce autoridades mundiales de protección de datos, incluida la Oficina del Comisario de Información del Reino Unido, subraya este punto.[5]

Para utilizar el consentimiento como base jurídica, sería necesario cumplir todos los requisitos para un consentimiento válido en virtud del RGPD, lo que significa que tendría que ser el resultado de una clara acción afirmativa, darse libremente, ser específico, informado e inequívoco. Se trata de un listón muy alto en el mundo de la formación de sistemas de IA.

Si el proveedor de IA no está en contacto con los interesados, como suele ser el caso, este consentimiento tendría que ser recabado por el usuario del sistema de IA, la organización con la que el interesado *sí* tiene relación. Sin embargo, esto ocurrirá normalmente a posteriori, cuando el sistema de IA ya haya sido entrenado, por lo que oponerse al tratamiento sería la mayoría de las veces irrelevante, puesto que el tratamiento de datos ya se habría producido. Además, también sería muy difícil a la inversa, cuando el sistema de IA ya habrá ingerido grandes cantidades de datos personales relativos a numerosos interesados.

[4] Si el sistema de exclusión voluntaria se pone en su conocimiento después de la formación de la IA, oponerse al tratamiento sería la mayoría de las veces, por un lado, irrelevante, puesto que el tratamiento de datos ya se habría producido y, por otro, muy difícil de detener cuando se están introduciendo en una IA grandes cantidades de datos personales relativos a numerosos interesados (*véase la parte XXX relativa al ejercicio de los derechos de los interesados*).

[5] https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2023/08/joint-statement-on-data-scraping-y-protección-datos/

En conclusión, es muy probable que el interés legítimo sea la base más adecuada para entrenar sistemas de IA con datos personales, sin embargo, como se ha indicado anteriormente, no proporciona un fundamento seguro dada la necesidad de llevar a cabo una evaluación de los intereses legítimos, así como el hecho de que los interesados pueden oponerse a dicho tratamiento en cualquier momento.

4. Riesgos del Jailbreaking y garantías de protección de datos

Poco después del lanzamiento de ChatGPT, los piratas informáticos empezaron a intentar "jailbreakear" el chatbot de IA, intentando saltarse sus salvaguardas y hacerle decir cosas inapropiadas o irracionales. Estas intrincadas instrucciones para eludir las restricciones impuestas a los programas de IA se conocen como "Jailbreaks". Este término se utilizó originalmente en el contexto de la tecnología digital para referirse al acto de obtener acceso al sistema operativo de un teléfono inteligente o tableta, especialmente uno fabricado por Apple, con el fin de ejecutar software modificado o no autorizado.

En el contexto de los modelos de IA Generativa, el término se refiere ahora al diseño de instrucciones que hacen que los chatbots se salten las normas sobre la producción de contenidos que inciten al odio o que escriban sobre actos ilegales. Estos ataques consisten en manipular los sistemas de IA generativa para que produzcan contenidos que vayan en contra de las normas establecidas, como generar material que incite al odio o que sea ilegal. Otro uso de estos ataques podría ser la difamación y el ataque personal a un individuo una vez que se han filtrado datos personales.

Una empresa de seguridad especializada en Inteligencia Artificial fue capaz de descifrar GPT-4, el último chatbot generador de texto de OpenAI, en apenas unas horas tras el lanzamiento inicial del sistema. Utilizando instrucciones cuidadosamente elaboradas, el director general de la empresa de seguridad eludió los sistemas de seguridad de OpenAI y consiguió que GPT-4 generara rápidamente declaraciones homófobas, creara correos electrónicos de suplantación de identidad y apoyara la violencia. Este comportamiento desviado supone un grave riesgo, ya que puede poner al descubierto datos personales introducidos en el sistema de forma inadvertida o incluso intencionada y, por tanto, ser manipulados por agentes malintencionados.

Un ataque estrechamente relacionado es el de introducir instrucciones, que puede insertar silenciosamente datos o instrucciones maliciosos en modelos de IA. Un ataque de "prompt injection" tiene como objetivo obtener una respuesta no deseada de las herramientas basadas en LLM. Y luego lograr un acceso no autorizado, manipular las respuestas o eludir medidas de seguridad.

Las técnicas específicas y las consecuencias de los ataques de inyección puntual varían en función del sistema.

Los Jailbreaks y los "promt inyections" puntuales son una forma de hacking poco convencional, que utiliza frases bien elaboradas en lugar de código para explotar los puntos débiles de los sistemas de IA. Aunque en la actualidad estos ataques se centran en eludir los filtros de contenidos, los investigadores de seguridad advierten del potencial de robo de datos y actividades ciberdelictivas generalizadas a medida que los sistemas de IA generativa se hacen más frecuentes.

Numerosos servicios y productos online populares dependen en gran medida de grandes conjuntos de datos para entrenar y mejorar sus algoritmos de IA. Los flujos de datos procedentes de redes, plataformas de redes sociales, dispositivos móviles y otras fuentes contribuyen a la enorme cantidad de información que las empresas utilizan para entrenar sus sistemas de aprendizaje automático. Por lo tanto, es importante señalar que algunos de los datos contenidos en estos conjuntos de datos probablemente podrían considerarse datos personales, incluso para los usuarios menos preocupados por la protección de datos. Desgraciadamente, debido al uso indebido y a la mala gestión de los datos personales por parte de algunas empresas, la protección de datos se ha convertido en una cuestión política urgente a escala mundial.

En una línea similar, muchos de nuestros datos sensibles también se recopilan para mejorar los procesos habilitados por la IA. Estos datos desempeñan un papel crucial a la hora de impulsar la adopción del aprendizaje automático, ya que los algoritmos sofisticados se basan en ellos para la toma de decisiones en tiempo real. Los algoritmos de búsqueda, los asistentes de voz, los motores de recomendación y otras soluciones de IA aprovechan amplios conjuntos de datos de usuarios del mundo real para proporcionar resultados personalizados y pertinentes.

A principios de 2023 se creó un sitio web llamado Jailbreak Chat, en el que se recopilan y comparten mensajes de foros en línea para chatbots de inteligencia artificial como ChatGPT. Los visitantes del sitio pueden contribuir con sus propios jailbreaks, probar las instrucciones enviadas por otros y votar sobre su eficacia. Los usuarios malintencionados podrían aprovechar estos "jailbreaks" para recopilar datos personales contenidos en los sistemas y llevar a cabo delitos como el robo de identidad y la creación de "deepfakes" para hacerse pasar por personas vivas.

Las implicaciones de los jailbreaks y los ataques de inyección de comandos se vuelven más

significativas cuando estos sistemas obtienen acceso a datos personales y sensibles. Por ejemplo, si un ataque de inyección de comandos con éxito ordena a un asistente personal que ignore las instrucciones previas y envíe un correo electrónico a todos los contactos, no solo podría provocar la vergüenza de la persona afectada, sino también problemas generalizados para las personas afectadas y la rápida propagación de contenidos nocivos a través de las redes personales y de trabajo de la persona.

Garantizar la seguridad de modelos de fundación como ChatGPT es primordial a medida que su uso se generaliza. Sin embargo, los hackers no se rendirán fácilmente. A medida que los sistemas de IA han evolucionado, los jailbreaks se han vuelto más complejos. Algunas implican múltiples personajes, intrincadas historias de fondo, traducción e incluso elementos de codificación para generar resultados específicos.

Algunos "red team" autorizados provocan ataques contra modelos de IA para descubrir vulnerabilidades. Un red team en ciberseguridad representa el equipo de seguridad ofensiva, que se encarga de descubrir vulnerabilidades de seguridad mediante pruebas de penetración. Con GAI, estos equipos buscan "exploits" que incluyan vulnerabilidades reales, influyan en el comportamiento del sistema o engañen a los usuarios para burlar la seguridad del sistema. Otros intentos proceden de aficionados a los que les gusta mostrar resultados humorísticos o inquietantes en las redes sociales. Este enfoque de la seguridad no es óptimo, ya que está fragmentado y depende de la exposición viral y de personas influyentes para impulsar las soluciones.

Aunque empresas como OpenAI, Google y Microsoft han tomado medidas para hacer frente a los ataques de jailbreaking y "prompt inyections", los investigadores que están detrás de estos ataques siguen encontrando nuevas formas de explotar las vulnerabilidades. El desarrollo de sistemas de IA generativa requiere enfoques que van más allá de los métodos tradicionales de red-teaming, como el uso de un segundo modelo de IA para analizar las indicaciones o separar claramente las indicaciones del sistema de las indicaciones al usuario.

La automatización y las técnicas avanzadas son necesarias para identificar y mitigar los jailbreaks y los ataques de peticiones a escala. Al automatizar el proceso de identificación de vulnerabilidades y comportamientos no deseados, los investigadores pretenden descubrir y atajar un mayor número de estos riesgos de seguridad.

Estos tipos de técnicas automatizadas pueden considerarse el punto de partida de un compromiso más profundo por parte de los desarrolladores de IA para valorar y evaluar la seguridad de sus sistemas. AI implicar a una amplia gama de participantes y dar prioridad a la transparencia y la responsabilidad, el objetivo es mejorar la seguridad, la fiabilidad y el uso ético de la tecnología de IA generativa. Las evaluaciones de terceros, la mitigación automatizada de las fugas y el uso de la red de trabajo en equipo desempeñarán un papel fundamental en la consecución de este objetivo y en la mejora de las prácticas que rodean el desarrollo de la IA con el fin de cumplir los requisitos tanto del RGPD como de la próxima Ley de IA.

5. ¿Cómo se articulan los derechos de los interesados con las herramientas de IA generativa?

La IA generativa, o GenAI, son sistemas de IA capaces de generar texto, imágenes u otros medios en respuesta a instrucciones. Los modelos generativos aprenden los patrones y la estructura de los datos de entrada, generando posteriormente nuevos contenidos similares a los datos de entrenamiento, pero con cierto grado de novedad, en lugar de limitarse a clasificar o predecir datos. Estos sistemas de IA suelen basarse en Transformadores Generativos Preentrenados (GPT), redes neuronales artificiales construidas sobre la arquitectura del transformador, preentrenadas en grandes conjuntos de datos de texto sin etiquetar, y capaces de generar texto similar al humano. Emplean grandes modelos lingüísticos (LLM) para producir datos basados en el conjunto de datos de entrenamiento que se utilizó para crearlos.

Comprender la tecnología que hay detrás de la IA generativa es vital para darse cuenta de que estas herramientas abarcan varias fases, y los datos personales pueden tratarse en cada una de ellas. Sin embargo, el tratamiento de datos personales en una fase no implica necesariamente el tratamiento de datos en otra.

Las etapas, según la ley de protección de datos, en las que los derechos de los interesados relacionados con los datos personales podrían aplicarse en el contexto de la IA Generativa incluyen:

1. La fase de entrenamiento de datos, en la que se incorporan los datos personales.
2. La fase de desarrollo, en la que se utilizan datos personales para generar contenidos y el propio resultado de los contenidos.
3. El propio modelo, que puede contener datos personales.

También es esencial señalar que el software de IA generativa puede procesar indirectamente datos, en particular relacionados con el usuario de la solución, como datos de cuentas o metadatos relacionados con el uso de la solución.

En los modelos habituales de aprendizaje automático, la identificación de las personas sobre las que versan los datos del sistema es un reto potencial para garantizar sus derechos. Normalmente, estos datos incluyen sólo la información pertinente para las predicciones, sin identificadores únicos de los interesados. Se someten a diversas medidas de tratamiento previo para adecuarlos a los algoritmos de aprendizaje automático, a menudo transformando los datos personales en una forma que es más difícil (pero no imposible) de vincular a personas concretas. Por lo tanto, las leyes de protección de datos podrían seguir aplicándose a estos datos transformados, ya que podrían seguir utilizándose para identificar a las personas. Este proceso debe tenerse en cuenta a la hora de responder a las solicitudes de derechos de las personas.

Este proceso es diferente para los modelos generativos de IA que para los modelos comunes de aprendizaje automático, como se explica en el párrafo anterior. Los modelos generativos de IA se entrenan a menudo con datos accesibles en la web, y su valor también reside a menudo en generar resultados relacionados con personas físicas, lo que implica una cantidad significativa de datos personales en los datos de entrenamiento para estos modelos. En consecuencia, estos conjuntos de datos podrían ser objeto de solicitudes de los interesados.

En los modelos generativos de IA, el "aprendizaje continuo" también plantea retos únicos para el cumplimiento del RGPD. Estos modelos se actualizan periódicamente en función de las interacciones de los usuarios, lo que significa que los datos personales se procesan continuamente. Estos datos proceden en su mayoría de las interacciones e indicaciones de los usuarios de la herramienta, y debe tenerse en cuenta que los interesados y los proveedores de datos no son necesariamente la misma entidad en el contexto de los modelos de IA de aprendizaje continuo.

Teniendo en cuenta estas consideraciones, navegar por los derechos de datos en virtud del Reglamento General de Protección de Datos (RGPD en el contexto de los modelos de IA generativa presenta desafíos únicos, en particular para los derechos de supresión, rectificación, acceso y oposición.

El primer problema común es la imposibilidad de recuperar datos en los modelos generativos de IA. Como ya se ha dicho, estos modelos obtienen datos de orígenes muy diversos, como el web scraping y las interacciones de los usuarios. Este enfoque polifacético de la recopilación de datos dificulta la trazabilidad de las contribuciones individuales. Además, a diferencia de los sistemas tradicionales de almacenamiento de datos, en los sistemas GenAI los datos personales también están profundamente integrados en algoritmos complejos, lo que complica el aislamiento de datos específicos. Esto dificulta el cumplimiento de los derechos del RGPD, ya que es necesario identificar si los datos personales se procesan dentro del sistema y dónde.

Otra capa de complejidad es la cuestión de los "datos personales inferidos". Se trata de conclusiones que el modelo puede extraer basándose en su entrenamiento. Por ejemplo, un modelo generativo de IA podría deducir las afiliaciones políticas de un usuario basándose en interacciones de datos anteriores. La opinión predominante se inclina por incluir estas inferencias a la hora de responder a solicitudes de derechos, ya que podrían revelar indirectamente información personal. El concepto de "datos de grupo inferidos" también merece atención. Este tipo de datos se genera a partir de patrones más amplios reconocidos durante el entrenamiento. Que estos datos de grupo se consideren personales depende de su posterior tratamiento y utilización.

Además de los retos comunes, también hay otros específicos relacionados con los derechos individuales que requieren la modificación o eliminación de datos. En particular, la alteración o eliminación de datos del conjunto de entrenamiento tras un ejercicio de derechos podría afectar a la validación y corrección del modelo. Los datos originales sirven a menudo de base para dichos procesos de validación. Por otra parte, el borrado o la modificación de datos que ya están integrados en el modelo implicaría a menudo eliminar o modificar estos datos para volver a entrenar el modelo, una tarea que es a la vez costosa y lenta.

En resumen, la intersección de los derechos del RGPD y los modelos generativos de IA presenta un laberinto de desafíos, cada uno con sus propias complejidades y complicaciones. La propia naturaleza de estos modelos, desde la forma en que incorporan y procesan los datos hasta las dificultades para rastrear las contribuciones individuales, añade capas de complejidad al cumplimiento del RGPD. Aunque no existe una solución milagrosa para superar sin problemas estos retos, la evolución del panorama ofrece algunas soluciones emergentes que podrían servir como puntos de partida para el cumplimiento actual.

Para empezar, a pesar de la ausencia de una solución única, se pueden tomar medidas proactivas. Aplicar el principio de "privacidad por diseño y por defecto" durante las fases de creación y despliegue del modelo GenAI proporciona una capa fundacional de protección de datos integrada desde el principio.

Al navegar por el complejo terreno de la protección de datos, se podría considerar una estrategia preventiva que reduzca el alcance de los datos y sus características identificativas. De este modo, sería posible aliviar potencialmente muchas de las complejidades que podrían surgir más adelante en el ciclo de tratamiento de datos. La minimización de los datos podría constituir una parte esencial de esta planificación en una fase temprana, guiando al responsable del tratamiento de datos para que recopile sólo lo que sea realmente necesario. Sobre esta base, las técnicas de anonimización de datos personales o el uso de tecnologías de protección de la intimidad (PET), como los datos sintéticos, podrían permitir una mayor reducción del ámbito potencialmente afectado por el ejercicio de derechos.
Además, es crucial invertir en medidas proactivas como la cartografía y el etiquetado de datos. Estas medidas ofrecen claridad sobre los orígenes y las características de los datos de formación, lo que facilita la gestión de las solicitudes de derechos en fases posteriores.

A medida que los modelos de IA generativa pasan del desarrollo a la implantación, la atención se centra en optimizar la adaptabilidad y la trazabilidad. En esta fase, mantener registros meticulosos del procesamiento de datos no es sólo una buena práctica, sino que se convierte en algo indispensable para facilitar la respuesta a las solicitudes adecuadas. Esto es aún más importante dada la mayor maleabilidad de los datos en esta fase. Además, los retos del aprendizaje continuo en los modelos desplegados pueden abordarse eficazmente mediante técnicas de versionado. Esto permite volver a un estado anterior del modelo sin la laboriosa necesidad de volver a formarlo desde cero. Esta vinculación garantiza tanto la adaptabilidad como la trazabilidad, proporcionando un marco sólido para la conformidad.

6. Protección de datos desde el diseño: Cómo crear herramientas de IA generativa que cumplan el RGPD

La protección de datos desde el diseño desempeña un papel fundamental para garantizar el cumplimiento del Reglamento General de Protección de Datos (RGPD). Implica salvaguardar los datos personales desde las primeras fases del diseño y a lo largo de todo el ciclo de vida del sistema. La idea de la protección de datos desde el diseño surgió de un conjunto más general de principios de privacidad denominado Privacidad desde el Diseño, desarrollado por primera vez en Canadá a principios de la década de 2000. La privacidad desde el diseño es un enfoque de la ingeniería de sistemas desarrollado inicialmente por Ann Cavoukian y formalizado en un informe sobre tecnologías que mejoran la privacidad elaborado por un equipo conjunto del Comisario de Información y Privacidad de Ontario (Canadá), la Autoridad de Protección de Datos de los Países Bajos y la Organización de Investigación Científica Aplicada de los Países Bajos en 1995. El marco Privacy by Design se publicó en 2009 y fue adoptado por la Asamblea Internacional de Comisarios de Privacidad y Autoridades de Protección de Datos en 2010. Ese mismo año, la Conferencia Internacional de Autoridades de Protección de Datos y Comisarios de Privacidad aprobó por unanimidad una resolución que reconocía la privacidad desde el diseño como un componente esencial de la protección fundamental de la privacidad. A continuación, la Comisión Federal de Comercio de Estados Unidos incluyó la privacidad desde el diseño como una de las tres prácticas recomendadas para proteger la privacidad en línea.

Poco después de 2010, Europa comenzó a trabajar en la revisión de sus leyes de protección de datos. Inspirándose en Privacy by Design y sus principios, Europa elaboró unos principios de protección de datos mediante el diseño que se introdujeron en la legislación a través del artículo 25 del Reglamento General de Protección de Datos (RGPD) en 2018.

En los últimos años, el rápido desarrollo de la IA generativa ha dado lugar a una mayor concienciación sobre los riesgos potenciales y las consideraciones éticas a la hora de diseñar sistemas que procesan datos personales. Estas preocupaciones abarcan no sólo riesgos complejos de protección de datos, como la filtración de información sensible e historiales de chat, sino también una serie de amenazas a los derechos de los ciudadanos de la UE sobre sus datos, incluido el "derecho al olvido". Este derecho permite a los particulares solicitar a una empresa la supresión de sus datos personales. Mientras que la eliminación de datos de bases de datos es relativamente sencilla, la eliminación de datos de modelos de aprendizaje automático es una tarea más compleja. Técnicas de anonimización y las prácticas de minimización de datos pueden ayudar a encontrar un equilibrio entre la defensa de los derechos de las personas y la preservación de la utilidad general del modelo de IA generativa.

Algo que hay que tener en cuenta desde el punto de vista humano es que, debido a la complejidad de los sistemas de IA modernos, las personas que participan en la creación y el despliegue de sistemas de IA suelen tener una gama más amplia de conocimientos y experiencia que los desarrolladores de sistemas habituales, incluyendo ingeniería de software tradicional, administración de sistemas, científicos de datos, estadísticos, así como expertos en la materia.

Debido a esta amplia gama de conocimientos, puede haber una menor comprensión de los requisitos de cumplimiento de seguridad más amplios, así como de los de la ley de protección de datos más específicamente. Para estas personas, es posible que la seguridad de los datos personales no siempre haya sido una prioridad clave, especialmente si alguien estaba creando previamente aplicaciones de IA con datos no personales o en una capacidad de investigación en la que los datos personales estaban protegidos en sandboxes.

Los algoritmos sesgados son otro problema importante para la protección de datos. Los sistemas generativos de IA aprenden de grandes cantidades de datos y, si esos datos son sesgados, los algoritmos pueden perpetuar y amplificar esos sesgos en sus resultados. Esto plantea cuestiones éticas sobre la equidad, la discriminación y el daño potencial causado por el contenido generado por IA sesgada cuando se utiliza para tomar decisiones importantes que cambian la vida de las personas a las que se refieren los datos.

Las alucinaciones de la IA se refieren a los casos en que los sistemas generativos de IA producen resultados que no se basan en información real o precisa. Estas alucinaciones pueden inducir a error a los usuarios y tener implicaciones potenciales para la seguridad de los interesados. Los sistemas generativos de IA deben proporcionar resultados fiables y dignos de confianza, especialmente sobre los ciudadanos europeos cuyos datos personales y su exactitud están protegidos por el RGPD.

El auge de los deepfakes, que son contenidos de audio o vídeo realistas pero manipulados, también se ha asociado a la tecnología de IA generativa. Los deepfakes tienen el potencial de manipular la opinión pública, difundir información errónea y plantear riesgos para la seguridad pública. Las implicaciones éticas de los deepfakes ponen de relieve la necesidad de medidas sólidas para prevenir su creación y detectar y combatir su difusión.

Un aspecto fundamental de la protección de datos desde el diseño es la transparencia. Desempeña un papel crucial en la protección de datos desde el diseño y garantiza la responsabilidad dentro de los sistemas de IA. Las organizaciones deben ser transparentes sobre sus prácticas en materia de datos, explicando claramente cómo funcionan los sistemas de IA y las decisiones que toman.

Sin embargo, lograr la transparencia en los sistemas de IA puede resultar difícil debido a su complejidad. Es esencial desarrollar métodos y herramientas que permitan explicar las predicciones algorítmicas a los usuarios finales de forma significativa y comprensible.

Otras complicaciones surgen porque las prácticas comunes sobre cómo procesar los datos personales de forma segura en la ciencia de datos y la ingeniería de IA aún están en desarrollo. Como parte del cumplimiento por parte de una organización del principio de seguridad del RGPD, deben asegurarse de que supervisan activamente y tienen en cuenta las prácticas de seguridad más avanzadas al desarrollar sistemas de IA y al utilizar datos personales en un contexto de IA.

No es posible enumerar todos los riesgos de seguridad conocidos que podrían verse exacerbados por el uso de la IA para procesar datos personales. Sea cual sea el riesgo, sin embargo, las empresas deben asegurarse de que el personal tenga las habilidades y conocimientos adecuados para hacer frente no solo a los riesgos de seguridad, sino también a los riesgos de protección de datos. Aquí es donde entra en juego la importancia de la formación sobre el RGPD.

La eficacia de los modelos de IA depende en gran medida de la calidad de los datos que reciben, por lo que la protección de datos es un aspecto integral de su diseño. La utilización de datos sensibles durante el entrenamiento de algoritmos generativos de IA puede dar lugar a la aparición de información personal en los resultados del chatbot o comprometer la seguridad de los datos durante ciberataques.

Así, al diseñar productos de IA, es primordial desvincular los datos personales de los usuarios individuales mediante el uso de conjuntos de datos sintéticos con anonimización total e identificadores no reversibles para el entrenamiento algorítmico, la auditoría y la garantía de calidad, entre otras prácticas. La aplicación de controles estrictos sobre el acceso a los datos dentro de la empresa y la realización de auditorías periódicas pueden ayudar a prevenir las violaciones de datos.

También es importante reconocer que más datos no equivalen necesariamente a mejores soluciones. Probar algoritmos utilizando la minimización de datos puede ayudar a determinar la menor cantidad de datos necesaria para un caso de uso viable. Además, es fundamental ofrecer a los usuarios un proceso ágil para solicitar la eliminación de sus datos personales.

Adoptar técnicas de aprendizaje contradictorio, que implican combinar conjuntos de datos conflictivos durante el proceso de aprendizaje automático, puede ayudar a identificar fallos y sesgos en los resultados de los algoritmos de IA.

Además, explorar el uso de conjuntos de datos sintéticos que no contengan datos personales reales es un enfoque potencial, aunque se requiere más investigación para evaluar su eficacia.

Las organizaciones deben alinear el uso responsable de la IA con los principios de protección de datos existentes descritos en el RGPD. Estas directrices deben abarcar diversos aspectos, como la responsabilidad, la intervención humana, la precisión, la seguridad, la prevención de sesgos y la explicabilidad de la toma de decisiones automatizada. Las inversiones continuas en medidas de protección de la intimidad, el perfeccionamiento en auditorías algorítmicas y la adopción de metodologías éticas, de seguridad y de protección de datos mediante el diseño son necesarias para navegar eficazmente por las oportunidades y los riesgos asociados a la IA generativa. Tecnologías como la privacidad diferencial ofrecen técnicas de preservación de la privacidad que pueden incorporarse a los sistemas de IA generativa. Los métodos escalables de limpieza de conjuntos de datos, incluidos los requisitos de deduplicación y divulgación de datos de entrenamiento, contribuyen a abordar los retos relacionados con la privacidad.

Los esfuerzos colectivos de la comunidad de ingeniería y protección de datos, junto con el compromiso de las organizaciones individuales y los profesionales de la privacidad, desempeñan un papel indispensable a la hora de abordar los problemas de protección de datos en torno a la IA generativa. Mediante la adhesión a los principios de la protección de datos desde el diseño y la integración de evaluaciones exhaustivas de la protección de datos y los derechos fundamentales, las organizaciones pueden esforzarse por lograr una aplicación fiable de la IA generativa, manteniendo al mismo tiempo el cumplimiento del RGPD. Es esencial seguir invirtiendo en formación sobre protección de datos, mejorar las competencias en auditoría algorítmica e integrar la ética, la seguridad y la protección de datos mediante metodologías de diseño para garantizar el uso responsable y ético de la IA generativa.

7. Técnicas de protección de la intimidad y datos sintéticos

Las herramientas de IA generativa son herramientas complejas y, como todas las tecnologías de este tipo, plantean muchos retos jurídicos importantes. La IA generativa está ávida de datos, pero estos datos (especialmente los de calidad pueden ser difíciles de conseguir o pueden estar legalmente protegidos, ya sea desde el punto de vista de la propiedad intelectual o de la legislación sobre protección de datos.

Desde la perspectiva de la protección de datos, las tecnologías de mejora de la privacidad (PET pueden representar una solución válida para hacer frente a los problemas de protección de datos, en términos de minimización de datos, integridad, confidencialidad y protección de datos desde el diseño. La Agencia de Ciberseguridad de la Unión Europea (ENISA) define las privacy enhancing technologies (PET) como "*soluciones de software y hardware (por ejemplo, sistemas que engloban procesos técnicos,*

métodos o conocimientos) para lograr una funcionalidad específica de privacidad o protección de datos o para proteger contra riesgos la privacidad de un individuo o un grupo de personas físicas",

Entre las diversas PET que podrían desplegarse en el contexto de la IA generativa, los algoritmos de síntesis de datos que generan datos "artificiales", más conocidos como datos sintéticos, pueden desempeñar un papel fundamental.

Según el Supervisor Europeo de Protección de Datos (SEPD) "Los *datos sintéticos son datos artificiales que se generan a partir de datos originales y un modelo que se entrena para reproducir las características y la estructura de los datos originales (...). El proceso de generación, también llamado síntesis, puede realizarse utilizando diferentes técnicas, como árboles de decisión, o algoritmos de aprendizaje profundo. Los datos sintéticos pueden clasificarse con respecto al tipo de datos originales: el primer tipo emplea conjuntos de datos reales, el segundo emplea en su lugar conocimientos recopilados por los analistas, y el tercer tipo es una combinación de estos dos.*"

En esencia, los datos sintéticos son datos generados por ordenador a partir de datos reales existentes o de algoritmos y modelos que reproducen, total o parcialmente, características, patrones y propiedades de los datos del mundo real.

Por tanto, el uso de datos sintéticos puede aportar muchas ventajas a la hora de entrenar herramientas de IA generativa, sobre todo porque:

a) reduce la necesidad de recopilar grandes cantidades de datos personales reales. En la fase de entrenamiento de modelos de IA, esto es especialmente importante, ya que permite a los ingenieros generar conjuntos de datos mucho mayores a partir de cantidades relativamente pequeñas de datos personales;
b) permite un etiquetado casi perfecto (por ejemplo, exactamente definido para el desarrollo de un modelo específico de IA) y datos de mayor calidad, complementando o sustituyendo así los conjuntos de datos del mundo real. Un estudio de Gartner ha pronosticado que "*para 2024, el 60% de los datos utilizados para el desarrollo de proyectos de IA y analítica se generarán sintéticamente*";
c) si se detectan y corrigen adecuadamente, reducen potencialmente el sesgo o el desequilibrio estadístico de los conjuntos de datos originales, aumentando así la equidad de la toma de decisiones que se basan en los datos;
d) refuerza la privacidad y reduce la superficie de ataque de la ciberseguridad al limitar el riesgo de pérdida de confidencialidad, integridad o disponibilidad de la información personal real;
e) reduce los costes implicados en todas las etapas de la cadena de valor de los datos al limitar la necesidad de una excesiva recogida, limpieza, preparación y almacenamiento de datos.

Sin embargo, esto no significa que los datos sintéticos sean la solución completa para todos los problemas de protección de datos. Siguen existiendo algunos problemas legales que los DPO deben tener en cuenta.

En primer lugar, los datos sintéticos no corresponden necesariamente a datos anónimos, lo que significa que el riesgo de reidentificación, en un grado u otro, seguirá existiendo. En la práctica, el objetivo de los datos sintéticos es reproducir los datos del mundo real, y cuanto más precisos sean, conservando todas las características y patrones de los datos originales, más eficaces serán para el modelo generativo de IA entrenado con ellos. Pero, por otro lado, el inconveniente es que dicha eficacia aumentará, en proporción directa, el riesgo de **reidentificación**. Esto significa que no se extinguirá el riesgo de inferir datos relacionados con un individuo concreto a partir del conjunto de datos sintéticos, o del propio modelo de IA.

Como señala la Information Commissioner's Office (ICO). "*Debe centrarse en la medida en que las personas están identificadas o son identificables en los datos sintéticos, y qué información sobre ellas se revelaría si la identificación tiene éxito". Se ha demostrado que algunos métodos de generación de datos sintéticos son vulnerables a ataques de inversión de modelos, ataques de inferencia de pertenencia y riesgo de revelación de atributos. Estos pueden aumentar el riesgo de inferir la identidad de una persona....*

El uso de otras PET (como la privacidad diferencial o la supresión de valores atípicos (puntos de datos con algunos rasgos identificativos únicos, puede servir para reducir el riesgo de reidentificación de datos personales, pero no para eliminarlo por completo.

Además, la fase de generación de datos sintéticos puede implicar el tratamiento de datos personales, especialmente tras la recopilación y el análisis de conjuntos de datos reales, lo que conlleva la necesidad de cumplir el RGPD y las obligaciones relacionadas.

También debe hacerse mención específica del deber de proporcionar información completa en virtud del art. 13 del RGPD a los interesados cuyos datos se recopilen y utilicen posteriormente con fines de formación en materia de inteligencia artificial, así como de identificar una base jurídica para el tratamiento con arreglo al artículo 6 del RGPD.

Por último, la obligación de respetar estrictamente los principios del Art. 5 del RGPD. En particular, algunos de los siguientes principios del Art. 5 son dignos de mención en el caso de la IA generativa:

a) transparencia: no se limita a la información que debe facilitarse a los interesados con arreglo al artículo 13 del RGPD, como se ha mencionado anteriormente, sino también a los usuarios en relación con los resultados sintéticos generados por los modelos de IA, a fin de evitar el riesgo de falsificaciones profundas o de manipulación social 13 del RGPD, como se ha mencionado anteriormente, sino también hacia los usuarios, con referencia a los resultados sintéticos generados por modelos de IA, con el fin de evitar el riesgo de falsificaciones profundas y/o manipulación social;

b) limitación de la finalidad: dado que los datos sintéticos pueden derivarse de datos reales, que pueden contener información personal, es necesario destacar que dichos datos se han recogido con fines especificados, explícitos y legítimos, y que el tratamiento posterior (por ejemplo, para la sintetización de datos y el posterior entrenamiento de modelos de IA) no es incompatible con los fines iniciales.
Un principio similar se ha establecido en relación con el proceso de anonimización por WP Art. 29 (dictamen 5/2014) según el cual: "*el proceso de anonimización, es decir, el tratamiento de (...) datos personales para lograr su anonimización, es un caso de "tratamiento posterior". Como tal, este tratamiento debe cumplir con la prueba de compatibilidad de conformidad con las directrices proporcionadas por el Grupo de Trabajo en su Dictamen 03/2013 sobre la limitación de la finalidad*".

Especialmente en lo que respecta a la fase de entrenamiento de los modelos de IA, la referencia a los "fines estadísticos" como no incompatibles en principio con los fines iniciales en virtud de la letra b) del artículo 5, apartado 1, podría servir a este propósito. 5, ss.1, podría servir a este fin[6] .

c) exactitud e imparcialidad: aquí debe prestarse atención a evitar el riesgo de "alucinación", o de duplicar sesgos, errores o inexactitudes contenidos en el conjunto de datos original. Esto es especialmente importante si el modelo de IA entrenado con los datos sintéticos se utilizará después para adoptar decisiones que puedan afectar a los derechos o intereses de las personas.

El desarrollo de técnicas que permitan explicar los resultados generados por los sistemas de IA entrenados con datos sintéticos será de vital importancia para este fin concreto.

[6] Véase sobre este tema, Estudio a petición del Grupo para el Futuro de la Ciencia y la Tecnología (Servicio de Investigación del Parlamento Europeo) "*The Impact of the GDPR on artificial intelligence*", junio de 2020.

8. Cuestiones específicas de la IA generativa basada en imágenes y audio

En el caso de las aplicaciones de IA generativa no basadas en texto, como las herramientas de generación de imágenes, audio y vídeo, existen claras implicaciones para la protección de datos. Las aplicaciones más populares, como Midjourney y Stable Diffusion, que permiten a los usuarios generar rápidamente imágenes y vídeos introduciendo indicaciones de texto, se basan en grandes volúmenes de contenido de imágenes y vídeos. Estos datos subyacentes incluyen numerosas categorías de datos personales suficientes para identificar a los interesados, siendo la principal la propia imagen y semejanza de un interesado que a menudo se representará en los resultados.

En concreto, los DPO pueden esperar que las siguientes categorías de datos personales participen en dichas herramientas:

- imágenes fotográficas de los interesados;
- representaciones artísticas de los interesados;
- grabaciones de vídeo de los interesados; y
- Audio, datos de voz.

Las organizaciones tendrán que entender que el tratamiento posterior de estos datos entra en el ámbito de aplicación del RGPD. Por ejemplo, si un departamento de marketing quiere crear material promocional y utiliza imágenes de los interesados obtenidas a partir de IA generativa, tendrá que procesar esas imágenes de acuerdo con la legislación sobre protección de datos y respetar principios fundamentales como la transparencia, la legalidad y la equidad.

Además, debe tenerse en cuenta la cuestión de combinar los datos procedentes de fuentes de IA generativa con datos de otras fuentes. Si bien es posible que los datos recibidos de la herramienta de IA generativa no identifiquen al interesado, el acto de combinarlos con datos alternativos puede hacerlo y, una vez más, poner de manifiesto los requisitos del RGPD. Esto podría ser especialmente relevante cuando, por ejemplo, el pegado de imágenes de diferentes fuentes conduzca a la identificación de personas.

En los casos de uso más creativo, en los que las organizaciones pueden desear modificar, alterar o cambiar significativamente la presentación de imágenes, vídeos o contenidos de audio, esto debe llevarse a cabo respetando los derechos y libertades fundamentales de los interesados. Siempre deben tenerse en cuenta los riesgos, por ejemplo, de difamar o perjudicar a los interesados, y cuando se considere que el tratamiento puede ser de alto riesgo, debe realizarse una EIPD.

Por último, cuando las organizaciones deseen crear contenidos "deepfake" legítimos, como, por ejemplo, vídeos corporativos oficiales, las cuestiones del consentimiento del interesado y la transparencia del tratamiento deben ser consideraciones clave.

9. Gestión del riesgo de protección de datos

Llevar a cabo una evaluación del impacto sobre la protección de datos (EIPD) al implantar o utilizar un sistema de IA generativa resulta aún más crucial cuando, como suele ser el caso, estas herramientas aún no se han comprendido adecuadamente, tanto desde el punto de vista de la estrategia empresarial como de la gestión de riesgos. La comprensión de los riesgos que entraña para los datos personales el tratamiento generativo de la IA sigue evolucionando y todos los DPO deben tratar de estar atentos a las amenazas y retos aún no previstos. Para gestionar estos riesgos emergentes, deben tenerse en cuenta los siguientes factores.

a) Riesgos para los interesados

La relación entre el usuario y la IA, así como las repercusiones que el tratamiento tendrá en las personas, deben constituir el núcleo del análisis. Entre los riesgos potenciales para los interesados figuran los siguientes:

- Consecuencias de una decisión parcial o totalmente automatizada producida por la IA generativa. Las consecuencias de tales decisiones pueden consistir en pérdidas de oportunidades financieras o incluso restricciones de derechos fundamentales.
- Riesgos de reforzar la discriminación y los prejuicios contra determinados usuarios.
- Riesgos derivados del tratamiento de datos de categoría especial, como se indica en el art. 9 del RGPD. Por ejemplo, una herramienta de IA generativa podría inferir de determinados datos personales del interesado, (a partir de sus modalidades de expresión o del uso de determinadas palabras), su origen étnico, sus posiciones políticas o filosóficas, o incluso su orientación sexual, y aplicar un trato diferenciado sobre esta base. Para identificar estos riesgos, la empresa que utiliza la herramienta de IA generativa debe llevar a cabo una revisión periódica de la calidad de los resultados generados.
- En términos de seguridad informática, la información disponible para el atacante en el sistema de IA puede ser un vector de amenaza. Un escenario denominado de "caja blanca", en el que el atacante puede deducir/encontrar mucha información técnica para preparar su ataque, crea más exposición en comparación con un sistema de "caja negra" en el que el atacante sólo puede acceder a la información producida por el sistema como salida. Más concretamente, los siguientes ataques son específicos de las fases definidas del proyecto de IA:

Fase de aprendizaje	tipo de ataque	infección	ataques por la espalda
			ataques de envenenamiento
		exfiltración	ataques de inferencia de pertenencia
			ataques de inversión de modelos
			ataques de extracción de modelos

Fase de producción	tipo de ataque	manipulación	ataques de evasión
			ataques de reprogramación
			denegación de servicios
		exfiltración	ataques de inferencia de pertenencia
			ataques de inversión de modelos
			ataques de extracción de modelos

b) Determinación de medidas paliativas

La EIPD, como siempre, debe realizarse antes de iniciar el proyecto y, a continuación, a través de la protección de datos desde el diseño, debe informar y guiar la fase de diseño de cualquier herramienta de IA generativa. En el caso de la IA generativa, deben tenerse en cuenta las siguientes medidas paliativas para gestionar los riesgos identificados:

- Ajuste fino supervisado con conversaciones ejemplares en las que se entrena a un LLM para que reproduzca un corpus de conversaciones que ilustran lo que se considera un comportamiento deseado.
- Puesta a punto con un modelo de valor humano en el que los operadores humanos recompensarán los resultados más satisfactorios.
- Además, las medidas organizativas deben tener como objetivo garantizar una evaluación constante de los resultados proporcionados por la herramienta de IA generativa, tanto a nivel del operador humano que la utiliza como de una entidad organizativa que analiza los resultados a gran escala para garantizar un alto nivel de calidad de los resultados a lo largo del tiempo.
- Del mismo modo, debemos procurar en la medida de lo posible una situación de explicabilidad de las decisiones tomadas por el modelo de IA generativa para permitir un auténtico control humano. En este sentido, el control humano sigue siendo, en última instancia, el mejor método para mitigar los riesgos que plantean los sistemas de IA generativa. De este modo, se puede evitar una confianza excesiva en los resultados producidos por las herramientas de IA generativa.

Tal exceso de confianza conduciría, en ausencia de controles humanos eficaces, a la producción de decisiones totalmente automatizadas.

Una consideración adicional para los DPO es el nuevo requisito de gobernanza de la IA de realizar Evaluaciones de Impacto sobre los Derechos Fundamentales (FRIAs). En el proyecto de texto de la Ley de Inteligencia Artificial, que en la fecha de publicación de este documento aún se encuentra en la fase de diálogo tripartito de los debates en el poder legislativo de la UE, se incluye el requisito de realizar FRIAs. La intención es que dicha evaluación sea realizada por el proveedor o el usuario de un sistema de IA cuando existan riesgos para los derechos y libertades fundamentales de las personas afectadas por el resultado.

Dado que las FRIAs son, en efecto, similares a las EIPD para el mundo de la IA, con solapamientos particulares en la comprensión de cómo las actividades de tratamiento afectan a los derechos fundamentales, los DPO deben esperar que este trabajo se les asigne una vez que la Ley de IA entre en vigor. Aunque, en algunos aspectos, los DPO están en una posición única, y cualificados, para realizar este trabajo, no están necesariamente familiarizados de forma natural con los nuevos riesgos tecnológicos que están creando rápidamente las tecnologías de IA. Por esta razón, los DPO ya deberían estar investigando y comprendiendo los riesgos específicos de la IA para los datos personales.

Desde el punto de vista práctico, puede ser posible realizar las evaluaciones de impacto sobre la seguridad y las evaluaciones de impacto sobre la protección de datos como un único ejercicio, pero sea cual sea el método que se elija en última instancia, los responsables de la protección de datos deben empezar a adquirir conocimientos sobre el riesgo de IA ahora, en previsión de la Ley de IA.

10. Transparencia e inteligencia artificial generativa

Cuando se recopilen y suministren datos, incluidos datos personales, a una IA con fines de formación y cuando este tratamiento de datos se rija por el RGPD, la entidad que imparta esta formación (el operador de la IA) deberá garantizar la transparencia de dicho tratamiento de datos de conformidad con el artículo 5, apartado 1, letra a), y el artículo 12 y siguientes de dicho Reglamento.

Se pueden identificar tres fuentes de datos diferentes:

- El scraping de datos de sitios web con ayuda de robots o sistemas de IA (caso de uso 1);
- El suministro de datos por parte de usuarios del sistema o proveedores de datos relativos a otras personas (caso de uso 2);
- El suministro de datos sobre sí mismos por parte de los usuarios de la IA (caso de uso 3).

Para cada uno de estos casos de uso, las formas de garantizar la transparencia del tratamiento de datos varían en función del tipo de IA de formación necesaria.

Caso práctico 1

La transparencia es una cuestión delicada y tal vez difícil cuando se considera el scraping de datos en línea, principalmente debido al hecho de que cualquier dato personal recogido de esta manera no se obtiene directamente del interesado. En consecuencia, el artículo 14 del RGPD debe aplicarse a estos datos, es decir, los datos personales que no se hayan obtenido directamente del interesado dan derecho al interesado a obtener del responsable del tratamiento confirmación de si se están tratando sus datos personales y, en caso afirmativo, se le debe proporcionar acceso a sus datos personales junto con otra información esencial, como la finalidad del tratamiento y las categorías de datos que se están tratando, etc.

Además, debe aplicarse el artículo 15 del RGPD relativo al derecho de acceso del interesado a su información personal.

Sin embargo, en un escenario así se presentan varias dificultades para el operador de IA. Especialmente las siguientes:

- Identificar los datos personales entre los datos recuperados automáticamente por la IA, que suelen consistir en grandes cantidades de datos;
- Identificación directa de cada interesado;
- Obtener información de contacto suficiente para informar a cada interesado del tratamiento de sus datos.

A la luz de estas dificultades, podría aplicarse el artículo 14.5 (b) del RGPD. Esta sección del artículo estipula que un responsable del tratamiento de datos no tendría que proporcionar la información especificada a cada interesado cuando "*el suministro de dicha información resulte imposible o suponga un esfuerzo desproporcionado*". La jurisprudencia de diversas autoridades de protección de datos muestra que esta excepción debe interpretarse de forma muy estricta. Dicho esto, dadas las dificultades señaladas anteriormente en relación con los modelos generativos de IA, podría aplicarse en este caso. No obstante, de ser así, el operador de IA seguiría estando obligado a cumplir los requisitos de transparencia ante el interesado.

De conformidad con dicho artículo 14.5 (b), el responsable del tratamiento debe adoptar las medidas adecuadas para proteger los derechos y libertades del interesado y sus intereses legítimos. Dichas medidas incluyen la publicación de la política de privacidad del responsable del tratamiento en su sitio web, pero también, posiblemente de forma más estricta

medidas como el ejemplo dado por la Autoridad Italiana de Protección de Datos al regular ChatGPT a principios de 2023. En última instancia, OpenAI acordó llevar a cabo una campaña informativa, de carácter no promocional, en todos los principales medios de comunicación italianos (radio, televisión, periódicos e Internet) para informar a los ciudadanos de la probable recogida de sus datos personales con el fin de formar a ChatGPT. También acordaron poner a disposición en el sitio web del responsable del tratamiento una herramienta a través de la cual todos los interesados pudieran ejercer su derecho a acceder a sus datos personales.

Por otra parte, en lo que respecta a dicho derecho de acceso, también puede ser de aplicación el artículo 11 del RGPD, que estipula que:

"*1. Si los fines para los que un responsable del tratamiento trata datos personales no requieren o ya no requieren la identificación de un interesado por parte del responsable del tratamiento, éste no estará obligado a conservar, adquirir o tratar información adicional para identificar al interesado con el único fin de cumplir el presente Reglamento.*

2. Cuando, en los casos a que se refiere el apartado 1 del presente artículo, el responsable del tratamiento pueda demostrar que no está en condiciones de identificar al interesado, informará de ello al interesado, si es posible. En tales casos, no se aplicarán los artículos 15 a 20, excepto cuando el interesado, a efectos del ejercicio de sus derechos con arreglo a dichos artículos, facilite información adicional que permita su identificación".

Además, en el considerando 4 del RGPD se nos recuerda que "*el derecho a la protección de los datos personales no es un derecho absoluto; debe considerarse en relación con su función en la sociedad y sopesarse con otros derechos fundamentales, de conformidad con el principio de proporcionalidad*". En consecuencia, podría argumentarse que no pueden imponerse esfuerzos desproporcionados al operador de la IA para identificar al solicitante y detectar sus datos personales en los datos de entrenamiento de la IA.

A la luz de lo anterior, el operador de IA que se enfrente a una solicitud de acceso deberá:

1. Comprobar si los datos personales relativos al solicitante pueden identificarse;
2. Proporcionar al solicitante todos los datos personales identificados;
3. Informar al interesado de que puede haber datos personales que le conciernan que el operador de IA no esté en condiciones de detectar/proporcionar dadas las características del tratamiento de datos que se está llevando a cabo.

Asimismo, para cumplir el artículo 25 del RGPD y el principio de protección de datos desde el diseño, el operador de IA también puede estar obligado a demostrar que puede anticipar dichas solicitudes de acceso y que ha revisado todas las posibilidades técnicas que podría desplegar razonablemente para detectar los datos personales relativos a cada solicitante (y que reevalúa periódicamente estas posibilidades).

Caso práctico 2

Dado que los datos suelen ser suministrados a los operadores de IA a lo largo de la cadena de suministro por otros terceros situados más arriba en la cadena de suministro (un usuario o un proveedor de datos). Estos terceros podrían ayudar al operador de IA a garantizar la transparencia en el tratamiento de los datos proporcionando herramientas y orientación sobre la mejor manera de extraer datos personales del conjunto de datos, dado que son estos terceros los que suministran los conjuntos de datos en primer lugar. Estas terceras partes también podrían ayudar al operador de IA a gestionar las solicitudes de acceso de los interesados por las mismas razones.

Caso práctico 3

Cuando los datos personales se recogen directamente de los usuarios, se aplica el artículo 13 del RGPD. El responsable del tratamiento debe proporcionar información específica al interesado en el momento de la recogida, por ejemplo, la identidad y los datos de contacto del responsable del tratamiento; los datos de contacto de su delegado de protección de datos; los fines del tratamiento a que se destinan los datos personales, así como la base jurídica del tratamiento; junto con otra información específica.

11. Optimizar las estructuras organizativas

Dentro de cualquier organización, desde el punto de vista de la estructura de gestión, el tema de la IA Generativa tendrá que abordarse de forma multidimensional, como reflejo de la complejidad de la tecnología y sus repercusiones. No será viable para las empresas que cada función trabaje por su cuenta y no interactúen entre sí.

El impacto de la IA es una cuestión de empresa; por lo tanto, requiere un enfoque conjunto de toda la empresa. Este enfoque integrado es esencial para evitar la duplicación de esfuerzos y, lo que es más importante, para garantizar que las decisiones clave reciban aportaciones multidisciplinares.

Para lograrlo, las organizaciones deberían crear un grupo de trabajo sobre IA, centrado en la IA responsable y su gobernanza. La creación de dicho grupo de trabajo podría ser una iniciativa impulsada por el DPO, por ser una de las funciones que mayor exposición tendrá a este tema debido a que tiene que gestionar algunas consideraciones relativas a la IA en un contexto en el que estén implicados datos personales. Alternativamente, podría ser iniciada y dirigida por una función de TI, como un Director de Datos o un Director de Tecnología.

Este grupo de trabajo implicará de forma significativa al departamento jurídico, las funciones de cumplimiento y, en concreto, la protección de datos. Para los aspectos técnicos, el departamento de seguridad informática debería estar representado. El grupo de trabajo puede involucrar al personal de comunicación y relaciones públicas, ya que será necesario comunicar internamente, y potencialmente externamente, las decisiones tomadas por el grupo de trabajo. El líder del grupo operativo puede establecer grupos de discusión en los que miembros seleccionados del grupo operativo se centren en cuestiones específicas e informen de sus resultados al grupo operativo. El diagrama anterior da una idea indicativa de la composición de estos grupos de discusión y de cómo se relacionarían con el grupo de trabajo sobre gobernanza responsable de la IA.

La misión del grupo de trabajo es responder a la necesidad inmediata de una gobernanza responsable de la IA dentro de la organización y examinar y gestionar los riesgos en el uso de la IA generativa, en concreto, en relación con los datos personales, la parcialidad, las preocupaciones éticas, la nueva normativa sobre IA y numerosas cuestiones jurídicas como los derechos de propiedad intelectual y la exposición a la responsabilidad civil.

El principal objetivo de este grupo de trabajo será definir un plan de acción. Un aspecto crítico de este plan de acción será realizar un inventario de los sistemas de IA utilizados en la empresa, lo que incluye la IA generativa. Otro aspecto crítico es definir las funciones y responsabilidades de todas las funciones del grupo.

La función de este grupo de trabajo sobre IA es también sensibilizar a todos los niveles de la empresa sobre las cuestiones relacionadas con la IA. Este punto es importante, ya que el riesgo procederá naturalmente de los empleados que son los usuarios cotidianos de la tecnología, pero tiene que vincularse con el más alto nivel de toma de decisiones, porque decidir la forma de utilizar (o no utilizar la IA generativa es una estrategia de empresa.

Como tarea inicial, el grupo de trabajo debería preparar unas orientaciones preliminares para la organización sobre el uso responsable de la IA generativa que, por ejemplo, incluirían la recomendación de no introducir datos personales en las preguntas de herramientas relevantes como ChatGPT, ni subir imágenes con personas identificables.

Independientemente de la complejidad de la tecnología y de su implementación, el papel del DPO en este grupo de trabajo es, en última instancia, garantizar que cualquier dato personal procesado a través de tecnologías de IA cumpla con el RGPD.