IUS ET SCIENTIA
2025 • Vol. 11 • Nº 2 • ISSN 2444-8478
https://editorial.us.es/es/revistas/ius-et-scientia
https://dx.doi.org/10.12795/IETSCIENTIA • © Editorial Universidad de Sevilla 2025


# La compatibilidad del web scraping con los principios de la protección de datos personales

## THE COMPATIBILITY OF WEB SCRAPING WITH THE PRINCIPLES OF PERSONAL DATA PROTECTION

**Pablo Agustín Viollier Bonvin**
Universidad Central de Chile
pabviobon@alum.us.es 0000-0001-9893-7974

RESUMEN

Este artículo analiza la compatibilidad del raspado web (web scraping) con la normativa europea de protección de datos personales, particularmente el Reglamento General de Protección de Datos (RGPD) y el Reglamento de Inteligencia Artificial (RIA) de la Unión Europea. A través de un estudio doctrinal y jurisprudencial, se examinan los principios fundamentales del tratamiento de datos y su tensión con el web scraping. Se evalúan los límites y excepciones aplicables al web scraping y el rol de la recolección de datos de fuentes públicas para el entrenamiento de sistemas de inteligencia artificial. Finalmente, se discuten los desafíos regulatorios y las brechas existentes en la normativa que requieren ser subsanadas mediante pronunciamientos interpretativos o a través de una solución regulatoria que garantice alcanzar un equilibrio entre la protección de los derechos de los titulares y el acceso de datos de entrenamiento para el desarrollo modelos de inteligencia artificial.

PALABRAS CLAVE

Protección de datos personales
Principios
Inteligencia artificial
Web scraping

ABSTRACT

This article analyzes the compatibility of web scraping with European personal data protection regulations, particularly the General Data Protection Regulation (GDPR) and the Artificial Intelligence Act (AI Act). Through doctrinal and jurisprudential study, the fundamental principles of data processing and their conflict with web scraping are examined. The limits and exceptions applicable to web scraping and the role of data collection from public sources for training artificial intelligence systems are evaluated. Finally, the regulatory challenges and existing gaps in the regulations that need to be addressed through interpretive pronouncements or a regulatory solution that guarantees a balance between the protection of data subjects' rights and access to training data for the development of artificial intelligence models are discussed.

KEYWORDS

Data protection
Principles
Artificial intelligence
Web scraping

IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04



## 1. Introducción

La necesidad de compatibilizar la normativa de protección de datos personales con el tratamiento que los sistemas de inteligencia artificial (en adelante "IA") realizan de este tipo de información ha sido ampliamente estudiado por la doctrina[1], la jurisprudencia[2] y se ha transformado en una importante preocupación por parte de los actores que participan del mercado de la tecnología[3]. Recientemente, con la dictación de la denominada "AI Act" de la Unión Europea[4], los legisladores también se han aventurado a regular esta materia.

Sin embargo, la etapa anterior al funcionamiento de los sistemas de inteligencia artificial, es decir, la recolección de información para el entrenamiento de los modelos o algoritmos no ha recibido la misma atención por parte de la literatura académica y los reguladores. Por regla general, los sistemas de inteligencia artificial requieren de gigantescas bases de datos para entrenar sus algoritmos, las que muchas veces son obtenidas de fuentes públicas como internet a través de un proceso denominado raspado web o "web scraping". Como ha señalado Almaqbali *et al.*, los datos son muy importantes para las empresas y organizaciones ya que ayudan en la toma de decisiones y, actualmente, la mayoría de los datos se pueden encontrar en Interne (Almaqbali *et al.*, 2019, p. 145). Esta importancia y dependencia de la recolección de datos en fuentes públicas de internet sólo se ha intensificado en el contexto del entrenamiento de sistemas de inteligencia artificial. Así, una reciente declaración pública más de 40 representantes de la industria digital han reclamado que el desarrollo de sistemas de IA generativa "[r]equiere normas claras, de aplicación coherente, que permita el uso de datos europeos" y que "[l]a toma de decisiones regulatorias se ha vuelto fragmentada e impredecible, mientras que las intervenciones de las Autoridades Europeas de Protección de Datos han generado una gran incertidumbre sobre qué tipos de datos pueden utilizarse para entrenar modelos de IA" (EUNeedsAI, 2024)[5].

Esta falta de certidumbre se produce por la colisión entre el tipo de tratamiento que realizan los desarrolladores de sistemas de inteligencia artificial –en particular al momento de recolectar grandes volúmenes de datos a través del web scraping de fuentes públicas en internet– y la forma en que el Reglamento General de Protección de Datos

1. Un metaanálisis sistemático de 73 publicaciones con revisión de pares respecto del debate en torno a la regulación de la IA se puede encontrar en Folberth, *et al.* (2022).

2. Previo a la entrada en vigencia del Reglamento de Inteligencia Artificial de la UE (en adelante el "RIA"), la principal fuente de jurisprudencia relativa a sistemas de inteligencia artificial y sistemas algorítmicos provino de la aplicación de las salvaguardas contenidas en el artículo 22 del Reglamento general de protección de datos personales de la Unión Europea (en adelante "RGPD"). Ver Medina (2022) y Viollier (2021).

3. Previo a la implementación de las primeras leyes de inteligencia artificial, el panorama y la discusión estuvo hegemonizada por iniciativas de autorregulación impulsadas por la industria, la que condujo procesos cuyo resultado fue una sumatoria de guías y pautas éticas en la forma de "soft law". Para muchos expertos, dichos marcos éticos eran insuficientes para enfrentar las posibles amenazas y daños asociados con el uso de tecnologías de inteligencia artificial (Regine, 2022). Ver, por ejemplo, Hagendorff (2020) y Jobin *et al.* (2019).

4. Una revisión del contenido de la nueva legislación europea, así como el proceso legislativo de la Unión Europea se puede encontrar en Akhtar (2023).

5. La traducción es mía.

de la Unión Europea (el adelante "RGPD") regula el requisito de que los responsables acrediten contar con una base de licitud para el tratamiento de datos personales, así como el cumplimiento de los principios del RGPD.

Sin embargo, como correctamente han identificado Solove y Hartzog (2024), la práctica del web scraping se encuentra en directa contraposición con todos y cada uno de los principios internacionalmente reconocidos relativos a la protección de datos personales. Del mismo modo, como correctamente ha señalado Hacker los aspectos jurídicos de la recolección y procesamiento datos de entrenamiento todavía representan, comparativamente, terra incognita en la literatura académica (Hacker, 2021, p. 259). Si bien el RGPD y el nuevo Reglamento de Inteligencia Artificial de la Unión Europea (en adelante "Reglamento de IA") han realizado un esfuerzo por regular el uso de algoritmos, todavía no se ha visto el mismo esfuerzo por regular la recolección de datos en el contexto del entrenamiento de modelos de IA. Por lo mismo, subsanar este vacío en la literatura resulta una condición necesaria para poder avanzar en muchas de las discusiones que se están dando actualmente respecto a la regulación de la IA.

A fin de aportar al debate esta investigación busca analizar la compatibilidad entre la práctica del web scraping, particularmente en el contexto de la recolección de grandes bases de datos para el entrenamiento de sistemas de IA, con los principios establecidos en el RGPD. El objetivo de la investigación es determinar si las autoridades de control en materia de protección de datos personales a nivel europeo cuentan con las herramientas legales para compatibilizar la práctica del web scraping con los principios contenidos en el RGPD o si, por el contrario, es necesaria otra fórmula regulatoria para compatibilizar la práctica del web scraping con las disposiciones del RGPD. Para ello, el trabajo se organizará en las siguientes secciones:



La primera sección realiza una revisión de literatura técnica y jurídica relativa a la práctica del web scraping a partir de cuatro elementos: i) la emergencia del web scraping durante la primera década de internet, ii) la evolución reciente de la práctica del web scraping, iii) el rol del web scraping en el desarrollo de sistemas de IA, iv) determinar si la recolección de datos a través de web scraping para el entrenamiento de modelos de IA constituye un tratamiento de datos personales conforme al RGPD, y v) las nuevas formas de exclusión del web scraping a través de mecanismos jurídicos (términos y condiciones) y técnicos (scripts de exclusión técnicos o robot.txt).

La segunda sección realiza un análisis de cada uno de los principios contenidos en el RGPD, a fin de analizar su compatibilidad con la práctica del web scraping. Se pondrá especial énfasis en los principios de minimización de datos y de limitación de la finalidad, puesto que son aquellos que tienen mayor impacto en la práctica del web scraping, en particular en el contexto de la recolección de grandes bases de datos de fuentes públicas en internet para el entrenamiento de sistemas de IA.

La tercera sección analiza el contenido del nuevo Reglamento de IA, a fin de determinar si este contiene alguna solución regulatoria respecto a la recolección de datos para el entrenamiento de modelos de IA. Para ello se estudiará especialmente el modelo de espacios controlados de prueba o "sandbox regulatorios".

La cuarta sección explora cuáles son las soluciones regulatorias que eventualmente podrían diseñarse e implementarse para subsanar la tensión entre la práctica del web scraping y el cumplimiento de los principios contenidos en el RGPD.

IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

Finalmente, se presentan conclusiones preliminares, así como la identificación de áreas en donde futuras investigaciones podrían profundizar en esta materia a fin de poder subsanar el vacío regulatorio identificado.

## 2. ¿Qué es el web scraping?

El diccionario de Cambridge define el raspado web o "web scraping" como "la actividad de tomar información de un sitio web o de la pantalla de una computadora y colocarla en un documento ordenado en una computadora"[6] . El glosario del sitio web especializado en esta materia denominado ScraperApi lo define como "un proceso automatizado de extracción de datos de sitios web mediante herramientas de software o scripts. Implica enviar solicitudes a servidores web, analizar la respuesta HTML o XML y extraer la información deseada mediante selectores o expresiones regulares"[7].

El proceso de web scraping consiste en tres etapas interrelacionadas: i) el análisis del sitio web, ii) raspado del sitio web o copia de la información del sitio y iii) organización de los datos (Milev, 2017). El web scraping difiere del minado de datos o "data mining" en el sentido de que el segundo involucra un análisis de los datos y muchas veces requiere de sofisticadas técnicas estadísticas (Krotov y Silva, 2018). De esta forma, el objetivo del web scraping como técnica es convertir datos web no estructurados en datos estructurados que se pueden almacenar y analizar en una base de datos o una hoja de cálculo central (Sirisuriya, 2015).

Si bien es posible realizar el proceso del web scraping de forma manual, es decir, que una persona natural simplemente copie y pegue manualmente el contenido de un sitio web usando su navegador personal, para alcanzar los niveles de eficiencia y escalabilidad necesarios para recolectar grandes volúmenes de datos el proceso necesariamente debe ser automatizado. Para ello se utilizan "scripts" o set de instrucciones que realizan una tarea automatizada, también denominado como "bots". En informática, un bot es un software autónomo que opera en una red (particularmente Internet) y puede interactuar con otros sistemas o usuarios (Khder, 2021, p. 156). Para alcanzar este objetivo generalmente se utilizan dos bots distintos: i) Un bot "araña" también conocido como rastreador o web-crawler que se utiliza para rastrear sitios web automáticamente agregarlos a un directorio o índice, y ii) Un bot raspador o web-scraper cuya misión es extraer datos de sitios web, enviando solicitudes a un sitio web y luego analizando el HTML u otros datos que se devuelven en la respuesta (Bhatia, 2016).

Sin embargo, el web scraping no ha cumplido la misma función durante toda la historia de internet, sino que ha ido adaptado a medida que la tecnología y su utilización también ha evolucionado.

6. Definición disponible en: https://dictionary.cambridge.org/dictionary/english/web-scraping (último acceso el 21 de marzo de 2025). La traducción es mía.

7. Definición disponible en: https://www.scraperapi.com/glossary/ (último acceso el 21 de marzo de 2025). La traducción es mía.



IUS ET SCIENTIA • 2025
Vol. 11 • N° 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

## 2.1. El rol del web scraping en los primeros años de internet

El año 1989 el científico británico Tim Berners-Lee creó lo que hoy conocemos como World Wide Web (www). Sólo dos años después Lee publicó el primer navegador, un sitio http:// albergado en su computador personal que buscaba darle acceso a quienes accedan a la www al resto de los sitios disponibles.

Sólo dos años después, el año 1993 Matthew Gray creó el primer concepto de un web crawler. Denominado "The Wanderer" el programa tenía como propósito medir el tamaño de la web, pero luego fue utilizado para generar un índice de sitios web denominado el "Wandex". Si bien su autor no lo diseñó para dicho propósito, esta sería la tecnología base para la creación de los primeros motores de búsqueda generales de internet[8]. El primer motor de búsqueda fue lanzado el mismo año y se denominó "JumpStation", utilizando la misma tecnología que hoy es la base para gigantes como Google, Yahoo y Bing, JumpStation se transformó en el primer motor de búsqueda basada en el web-crawling y logró indexar millones de sitios web.

El motor general de búsqueda de Google fue oficialmente lanzado al mercado el 4 de septiembre de 1998 utilizando un algoritmo de búsqueda más eficiente llamado "BackRub", pero sobre la base de la misma tecnología de web-crawling y web-scraping desarrollada en el pasado. Si bien la principal labor del motor de búsqueda de Google es realizar web-crawling de internet para indexar los distintos sitios web y generar un índice de búsqueda a través de su "Googlebot", dicha labor necesariamente involucra cierto nivel de "raspado" o web-scraping. Como Google señala en su mismo sitio web "Google descarga texto, imágenes y vídeos de las páginas que encuentra en Internet con programas automatizados llamados crawler"[9].

Por otro lado, iniciativas altruistas y que buscan promover el interés público también dependen de la realización de web scraping. Así, por ejemplo, el historiador Jason Scott se embarcó en una iniciativa de preservación de la historia de internet. Preocupado por la eliminación de grandes sitios web que cuyo contenido se perdería para siempre, Scott fundó el Archive Team en 1996[10] y desde entonces la iniciativa ha utilizado web-crawling y web-scraping para respaldar todos los sitios de internet posibles[11], para así preservarlos y proporcionar nuevas oportunidades para que los académicos analicen estas plataformas ahora desaparecidas (Sellars, 2018, p. 373).

Cómo es posible apreciar, el web-crawling y web-scraping no sólo son la base de iniciativas que buscan promover el interés público, sino que también forman la base técnica respecto de cómo opera internet en los términos que hoy conocemos, toda vez



IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

8. Una breve relación de estos eventos se encuentra en: https://webscraper.io/blog/brief-history-of-web-scraping (último acceso el 10 de noviembre de 2025).

9. Información disponible en: https://developers.google.com/search/docs/fundamentals/how-search-works (último acceso el 21 de marzo de 2025). La traducción es mía.

10. Su sitio se encuentra disponible en: www.internetarchive.org (último acceso el 21 de marzo de 2025).

11. Si el lector siente curiosidad respecto a cómo se veían sus sitios web favoritos en el pasado, la herramienta "Wayback Machine" permite acceder a los respaldos periódicos realizados por el Internet Archive. Disponible en: https://web.archive.org/ (último acceso el 21 de marzo de 2025).

que los motores de búsqueda dependen de esta actividad para poder realizar su labor en el ecosistema de internet.

## 2.2. Web-scraping en la era de la web 2.0

El término web 2.0 es generalmente atribuido a Tim O'Reilly que lo escribió como un "punto de inflexión para la web" (O'Reilly, 2007, p. 17) donde las empresas que sobrevivieron al colapso de las puntocom comenzaron a usar la web como plataforma, en lugar de crear productos y servicios para su uso como clientes de escritorio (Wilson *et al.*, 2011, p. 2). El nacimiento de redes sociales como Facebook y LinkedIn permitieron a los usuarios mantener conexiones con fines sociales y profesionales. La emergencia de plataformas como YouTube, Flickr y blogs generados por los usuarios permitieron la creación y el intercambio de contenido a nivel mundial, y su facilidad de uso impulsa una "revolución de la contribución" (Cook, 2008, p. 60). Esta era de la web estuvo dominada por la creación de contenido por parte de los usuarios, en contraposición del contenido creado por los propios administradores de las plataformas, un fenómeno que llamó a algunos expertos a aprovechar la inteligencia colectiva de los usuarios y crear servicios que mejoren cuantas más personas los utilicen (O'Reilly y Battele, 2009, p. 1).

Este período, que comenzó cerca del año 2004 hasta nuestros días ha estado dominado por el uso del web scraping para la obtención valiosa de información para distintos usos, tales como: i) investigación científica, ii) precios y dinámicas de mercado, iii) tendencias predominantes en la web, iv) prácticas empleadas por la competencia, v) hacer más eficiente el funcionamiento del comercio electrónico, entre otras (Henrys, 2021, p. 1).

En otras palabras, en la era de la web alimentada por el contenido creado por los usuarios, el web scraping ha servido como mecanismo de recolección de inteligencia sobre tendencias de mercado, comportamiento de los usuarios y otras tendencias que han permitido a las organizaciones tomar decisiones basadas en evidencia para hacer más eficientes y personalizados sus productos y servicios.

## 2.3. Web scraping en la era de la inteligencia artificial

Como han señalado Solove y Hartzog;

> Los sistemas de inteligencia artificial dependen de cantidades masivas de datos, que a menudo se obtienen mediante el método de "scraping", es decir, la extracción automática de grandes cantidades de datos de Internet. El scraping permite a los actores recopilar enormes cantidades de datos personales de forma económica y rápida, sin previo aviso, consentimiento u oportunidad de oposición o exclusión voluntaria por parte del interesado de los datos (Solove y Hartzog, 2024, p. 4)).

Por regla general –y dependiendo del tipo de sistema que se quiera entrar– mientras mayor es la cantidad de datos que se alimentan al entrenamiento del sistema de IA, mejor es el resultado del modelo, más precisas sus predicciones y los resultados (outputs) de mejor resultad (Bowles *et al.*, 2018).



IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

El desarrollo IA requiere volúmenes masivos de datos para entrenar sus algoritmos, especialmente cuando se trata de las técnicas más novedosas y sofisticadas. Así, por ejemplo, "las aplicaciones de aprendizaje automático (Machine Learning) utilizan volúmenes de datos excepcionalmente grandes, que son analizados por una aplicación de aprendizaje automático para determinar las interrelaciones entre estos datos" (Tschider, 2021, p. 107). En el caso de la IA generativa, el Information Commissioner´s Office (ICO) del Reino Unido, en un proceso de consulta pública sobre bases de licitud para el web scraping para entrenar modelos de IA, señaló que a su entender al día de hoy la mayoría de los sistemas de IA generativa sólo pueden ser entrenados a partir de web scraping de gran escala (ICO, 2024a). De acuerdo con los hallazgos del ICO, al día de hoy existe poca evidencia que los modelos de IA generativa podrían ser desarrollados a partir de bases de datos más pequeñas y propietarias.

La recolección intensiva de datos contenidos en fuentes públicas de internet ha llegado a un punto tal, que los desarrolladores de sistemas de IA están enfrentando un desafío paradójico: a medida que herramientas como Stable Diffusion generan imágenes basadas en inteligencia artificial y GPT–4 crea texto, este contenido es vuelto a publicar en internet, el que nuevamente es recolectado a través del web scraping y utilizado para entrenar los algoritmos de inteligencia artificial. Esto significa que el out-put generado a partir de este entrenamiento incorpora la generación de datos generados recursivamente (recursively generated data), fenómeno que incluso puede implicar el colapso de los modelos de aprendizaje (Shumailov *et al.*, 2024, p. 759). A este fenómeno también se le ha denominado Desorden de Autofagia de Modelos o "MAD" por sus siglas en inglés (Alemohammad *et al.*, 2023).



Por otro lado, este modelo de recolección indiscriminada de datos ha producido una avalancha de acciones legales, Así, por ejemplo, una acción de clases ingresada en San Francisco (PM v. OpenAI LP) invoca entre otros elementos el robó información privada, incluida información de identificación personal, de cientos de millones de usuarios de Internet, incluidos niños de todas las edades, sin su conocimiento informado o consentimiento. La demanda invoca infracciones a la Electronic Communications Privacy Act, la Computer Fraud and Abuse Act y la California Invasion of Privacy Act de EEUU (IAPP, 2023).

De esta forma, el rol del web scraping en la era de la inteligencia artificial es el de recolectar de forma masiva e indiscriminada volúmenes ingentes de información disponible públicamente en internet para entrenar los algoritmos de los sistemas de inteligencia artificial. Esto, a su vez, ha generado importantes fricciones con distintos cuerpos normativos, elemento que explicaré en la sección 3 de este escrito.

## 2.4. ¿Es el web scraping una forma de tratamiento de datos personales en los términos del RGPD?

En esta materia vale la pena realizar un paralelo con un problema similar respecto al respeto de los derechos de propiedad intelectual. Al recolectar grandes volúmenes de datos de internet para el entrenamiento de modelos de IA, los desarrolladores –incluso sin intención– recolectan y utilizan información sujeta a protección por figuras de propiedad intelectual (OECD, 2025). Al momento de recolectar y utilizar contenido pro-

IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

tegido por propiedad intelectual, el desarrollador deberá demostrar que cuenta con la autorización del titular de los derechos o, en su defecto, que se encuentra amparado por una excepción en la legislación de propiedad intelectual. Esto pone a los desarrolladores en una situación difícil, debido a la dificultad de escalar la obtención de la autorización de los distintos titulares de derechos en la información recolectada. Por otro lado, existe una delgada línea entre los usos permitidos a través de excepciones como los usos justos o "fair use" e incurrir en una infracción, situación que debe generalmente ser analizada caso a caso (Jayachandran y Arni, 2023, p. 15)[12].

Algo similar sucede en materia de protección de datos personales. Al recolectar enormes volúmenes de datos de fuentes de acceso público, necesariamente dentro de dicha información un determinado porcentaje estará compuesto por información que constituya datos personales, datos personales sensibles o categorías especiales de datos de acuerdo con el RGPD[13]. Por tanto, es necesario discernir si la recolección y procesamiento de datos personales en el contexto del entrenamiento de modelos de IA constituye un tratamiento de datos conforme al RGPD. Si la respuesta es negativa, entonces los desarrolladores podrán utilizar libremente esta información. Sin embargo, si se considera que el web scraping implica un tratamiento de datos personales, entonces los desarrolladores deberán poder acreditar que se encuentran amparados por una base legal que otorgue licitud a dicho tratamiento, así como el cumplimiento de los principios y demás requisitos contenidos en el RGPD.

Al respecto, vale la pena seguir lo desarrollado por Li *et al.* en su excelente trabajo publicado a comienzos del 2025. Los autores recalcan que la naturaleza particular de los modelos de IA generativa, al operar en base a grandes bases de datos, las que son tokenizadas y tratadas de manera agregada, tienen el potencial de ofuscar el vínculo directo entre la información identificable de los individuos, al mismo tiempo que potencialmente generan contenido (outputs) basados o que reproduzcan de dicha información identificable (Li *et al.*, 2025, p.3). Por lo mismo, discernir si los modelos de IA tratan datos personales de forma intencional sigue siendo un tema objeto de debate (Moerel y Storm, 2024).



En particular, se podría argumentar que la tokenización[14] de la información y su utilización de modo agregado y a gran escala implica que el vínculo entre el interesado

12. Por ejemplo, un modelo de lenguaje puede utilizar las entradas de distintos blogs disponibles públicamente en internet para entrenar su modelo de IA (Large Language Model o LLM). Sin embargo, esas entradas de blogs son de titularidad de sus autores y su utilización puede constituir una infracción de sus derechos de propiedad intelectual si el desarrollador no es capaz de demostrar estar amparado en una excepción al requisito de contar con su autorización. Un buen resumen de este debate puede encontrarse en Guadamuz (2024).

13. El RGPD define, en su artículo 4 n° 2 tratamiento como "cualquier operación o conjunto de operaciones realizadas sobre datos personales o conjuntos de datos personales, ya sea por procedimientos automatizados o no, como la recogida, registro, organización, estructuración, conservación, adaptación o modificación, extracción, consulta, utilización, comunicación por transmisión, difusión o cualquier otra forma de habilitación de acceso, cotejo o interconexión, limitación, supresión o destrucción". Cómo es posible verificar, el listado menciona explícitamente la recogida de datos personales y aclara que esta puede ser por procedimientos automatizados o no.

14. El diccionario de Cambridge define tokenización como "el proceso de reemplazar una pieza privada de datos con un token (= una pieza diferente de datos que representa la primera), para evitar que la información privada sea vista por alguien que no tiene permiso para hacerlo".

IUS ET SCIENTIA • 2025
Vol. 11 • N° 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

y la información es ofuscado a un punto tal que la información no es semánticamente discernible para los seres humanos (Neel, 2024). En base a esta situación, algunos argumentan que los modelos de IA no tratan datos personales en un sentido relevante del término, por lo que no caería bajo la aplicación del RGPD (Li *et al.*, 2025, p.13). Esta postura defiende que el procesamiento de datos es "incidental" o "agnóstico" respecto a la existencia de datos personales por lo que estaría excluido de la aplicación del RGPD. Así, por ejemplo, el Comisionado para la Protección de Datos y la Libertad de Información de Hamburgo publicó un artículo sobre grandes modelos de lenguaje (large language models o LLMs) en donde argumentó que los LLM no procesan datos personales puesto que i) Los tokens sólo son representaciones matemáticas abstractas y ponderaciones de probabilidad, las que carecen de información individual, ii) El resultado de los LLM es de carácter probabilístico y no una reproducción del contenido y iii) El riesgo de extracción de datos personales del modelo de LLM ya entrenado no es relevante (HmbBfDI, 2024).

Sin embargo, como muestran Li *et al.* la auditoría de bases de datos populares (datasets) recolectados a través de web scraping sugiere que dichos repositorios frecuentemente incluyen datos personales recolectados indiscriminadamente desde internet (Li *et al.*, 2025, p.3). Así, la auditoría de bases de datos como LAION-500M (Birhane, *et al.*, 2023) o Common Crawl/C4 (Baack, 2024) muestra que estos datasets contienen información personal e información personal sensible, lo que vuelve difícil argumentar que la recolección de dicha información no constituye un tratamiento de datos personales o queda fuera de la aplicación de las disposiciones del RGPD.

A este respecto, concuerdo con Li *et al.* en el sentido que la naturaleza del web scraping a escala masiva de sitios web hace que el involucramiento de datos personales se vuelva algo casi inevitable, por lo que la postura que defiende que la recolección de datos a través de web scraping no constituye un tratamiento de datos personales bajo el GDPR se transforma en una postura radical y difícil de defender (Li *et al.*, 2025, p.16). Adicionalmente, a mi parecer existe una confusión en las distintas etapas de tratamiento. Puede ser que al momento de alimentar y entrenar el modelo de IA los datos personales hayan sido tokenizados o anonimizados, sin embargo, para ser anonimizados estos datos deben haber sido recolectados en su estado original. El mero hecho de recolectar los datos personales en su estado original implica un tratamiento de datos personales que debe estar respaldado por una base de licitud y el cumplimiento de los principios del RGPD, independiente que posteriormente los datos sean anonimizados u ofuscados. El hecho que los datos recolectados masivamente incluyan datos de carácter sensible o categorías especiales de datos complica aún más la situación, puesto que el RGPD establece medidas más restrictivas para su tratamiento. Adicionalmente, incluso la tecnología más avanzada no es capaz de excluir este tipo de datos de las bases de datos de entrenamiento. (Li *et al.*, 2025, p. 15).

El ICO del Reino Unido ha sido categórico en desechar la tesis del procesamiento incidental o agnóstico de datos personales en el contexto del web scraping. En su respuesta a los ingresos en el proceso de consulta sobre IA generativa el ICO aclaró que, si bien muchos desarrolladores de modelos de IA expusieron en el proceso de consulta pública que ellos no buscan o procesan datos personales intencionalmente y que dicho procesamiento es puramente incidental, la postura del ICO es clara: "la regulación de



IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

protección de datos personales aplica al procesamiento de datos personales (lo que incluye categorías especiales de datos), sin distinguir si dicho procesamiento es "incidental" o no intencional" (ICO, 2024b, p. 6).

## 2.5. Nuevas formas de exclusión: políticas legales y scripts de exclusión técnicos

En respuesta a la tendencia de recolección masiva de datos, muchos sitios web han optado por proteger su contenido del scraping web. Esta reacción puede deberse a que las actividades de los web-crawlers o scrapers sobrecargan la capacidad de navegación o porque los administradores del sitio quieren proteger el contenido alojado en este de la recolección masiva que implica el scraping web, ya sea porque se trata de información protegida por propiedad intelectual, información de carácter personal o información comercialmente relevante.

Esta exclusión puede tomar dos formas: una técnica y otra jurídica. El mecanismo técnico es denominado robot.txt el cual es un archivo de texto utilizado para implementar el protocolo de exclusión de robots o Robots Exclusion Protocol (REP). El uso de archivos robot.txt a través del Robots Exclusion Protocol ha aumentado significativamente con el tiempo, sin embargo, no se trata de una barrera técnica de carácter infranqueable. Por lo mismo, web crawlers éticos y la mayoría de las grandes empresas comerciales respetan lo establecido en el archivo de exclusión, sin embargo, otros actores pueden eludir dicha exclusión (Yang *et al.*, 2007).



Del mismo modo, es cada vez más común que los términos y condiciones de los sitios web establezcan como condición para la utilización del sitio web que la información contenida en este no sea recolectada de forma automatizada. De esta forma, se excluye jurídicamente la práctica del web scraping.

La negativa de los sitios web para permitir el web scraping de su contenido ha dado pie a lo que algunos expertos han denominado una "crisis del consentimiento" que ha generado un rápido declive en la información disponible para entrenar sistemas de inteligencia artificial. Así, el estudio realizado por Longre *et al.* concluyó que:

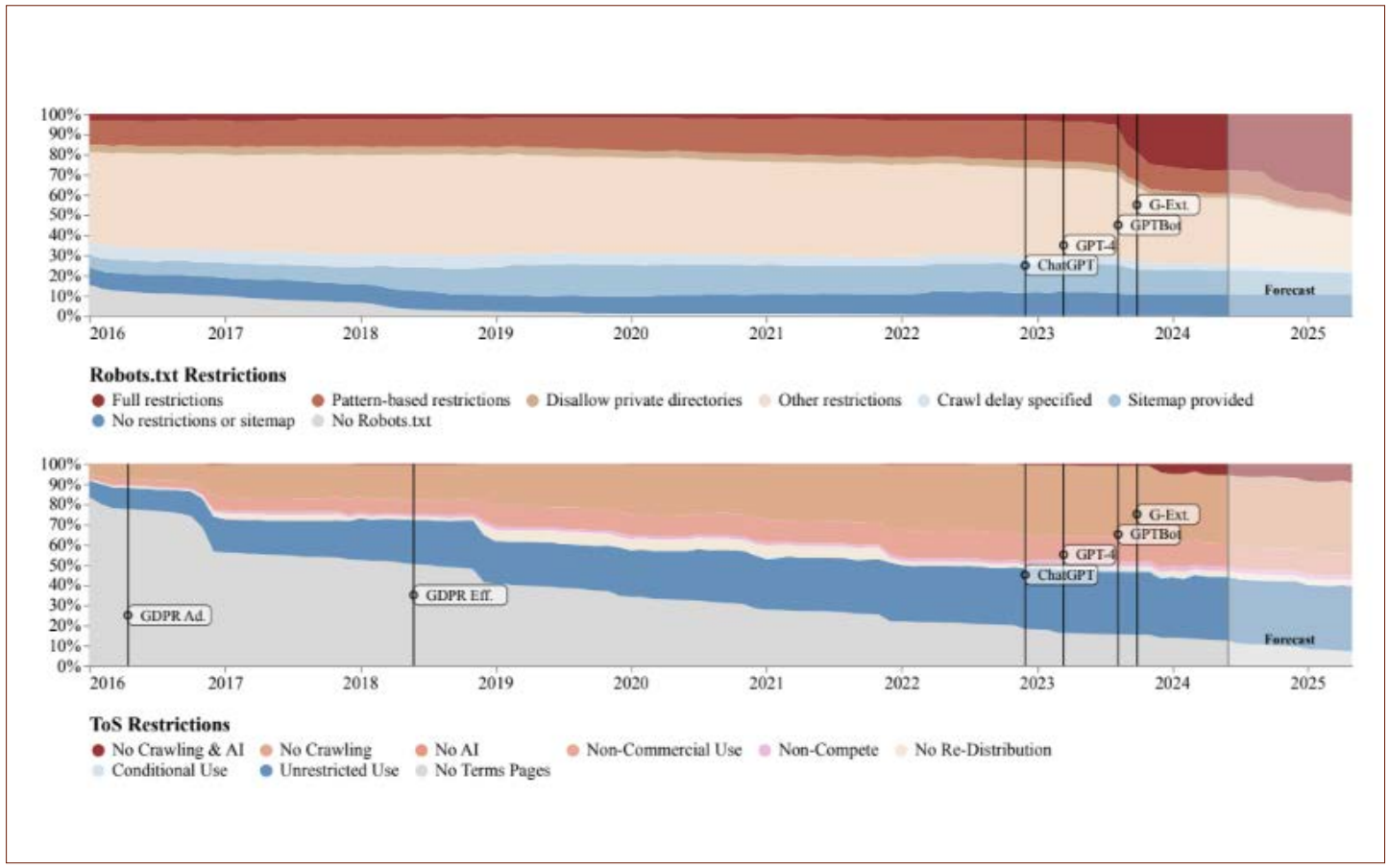
> El análisis realizado muestra una disminución clara y sistemática del consentimiento para rastrear y entrenar datos en la web. En la medida en que se respete este consentimiento, también se predice una disminución de los datos abiertos, lo que puede afectar a más personas que a los desarrolladores de IA comerciales o incluso a las organizaciones de IA en general (Longre *et al.* 2024, p.4).

Esta afirmación puede verse reflejada en la Figura 1.

La Figura 1 muestra que, a partir de 2016, la proporción de sitios que utilizan exclusiones de carácter técnicas (robots.txt) o jurídicas (términos y condiciones) para evitar el web scraping ha aumentado significativamente, esto refleja una adopción emergente de estas prácticas para excluir el web scraping y proteger el contenido de los sitios web.

IUS ET SCIENTIA • 2025
Vol. 11 • N° 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

**Figura 1**




*Fuente:* (Longpre *et al.* 2024, p. 5).

Sin embargo, el cumplimiento del contenido del archivo robot.txt dependen del cumplimiento voluntario por parte de los actores y muchas veces la observancia de los términos y condiciones de un sitio web son difíciles de aplicar jurídicamente. De esta forma, muchos sitios web han denunciado que, a pesar de tomar estas precauciones para excluir esta práctica, siguen siendo objeto de web scraping agresivo por parte desarrolladores de IA, incluso afectando el funcionamiento técnico de sus sitios web. En julio de 2024 varios sitios de noticias acusaron a la start-up Anthropic de llevar el nivel de recolección de datos a un nivel exagerado. El administrador del sitio Freelancer.com acusó que recibieron 3.5 millones de visitas del crawler de Anthropic en un período de tres horas, declarando que "[l]os motores de búsqueda siempre han realizado actividades de scraping, pero con el entrenamiento de la IA generativa el volumen de scraping ha llegado a otro nivel" (Financia Times, 2024).

Esta crisis del consentimiento repercute en el desarrollo de modelos de IA de dos formas distintas. El carácter de observancia de las medidas de exclusión abre el paso para que sean las organizaciones éticas y que cumplen la legislación las que queden privadas de los datos necesarios para desarrollar sistemas de IA, mientras que las organizaciones que deciden ignorar dichos resguardos todavía tendrán acceso a dichos datos. Por el otro, la disminución en la información disponible para entrenar sistemas de IA puede agravar la concentración de mercado y la posición dominante de los grandes gigantes de internet, quienes ya cuentan con un alto volumen de clientes y de grandes bases de datos que pueden ser utilizadas para entrenar sus propios sistemas de inteligencia artificial.



IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

## III. ¿Es compatible el web scraping con los principios contenidos en el RGPD?

Habiendo introducido el concepto de web scraping y sus aspectos técnicos, esta sección se aboca al objetivo principal de este trabajo: discernir si el web scraping para la recolección masiva de datos para el entrenamiento de sistemas de inteligencia artificial es compatible con los principios rectores del RGPD.

Este ejercicio ya ha sido realizado por Solove y Hartzog, quienes concluyeron que el web scraping es contrario a los principios de la protección de datos personales. En sus palabras:

> El scraping de datos personales viola casi todos los principios clave de privacidad incorporados en leyes, marcos y códigos de privacidad, incluidos la transparencia, la limitación de finalidad, la minimización de datos, la elección, el acceso, la eliminación, la portabilidad y la protección. El scraping implica la extracción masiva y no autorizada de datos personales para fines no especificados sin ninguna limitación o protección. En casi todas las dimensiones, esta práctica es antitética a la privacidad (Solove y Hartzog, 2024, p.4).

Sin embargo, dicho ejercicio fue realizado contrastando la práctica con la legislación estadounidense. La novedad de este trabajo yace en el contraste entre la práctica del web scraping con los principios específicos del GDPR.

### 3.1. Los principios en el RGPD



Los principios cumplen principalmente cuatro funciones respecto a la legislación de protección de datos personales: i) una función orientadora: guiando a los reguladores en la aplicación e interpretación del derecho, ii) una función integradora: permitiendo la armonización de distintas normas que pueden parecer fragmentadas o incluso contradictorias, iii) una función protectora: refuerzan la protección de los derechos del interesado, otorgando un piso mínimo incluso en aquellos casos en que la legislación no ha regulado específicamente, y iv) una función de flexibilidad normativa: permiten que la legislación se adecue a los cambios tecnológicos y a la emergencia de tecnologías disruptivas sin necesidad de estar constantemente modificando las normas frente a nuevas realidades.

El RGPD consagra explícitamente en su artículo 5 un catálogo de principios relativos al tratamiento, los cuales fueron debidamente definidos y dotados de fuerza normativa. Por otro lado, estos principios no sólo sirven de sustento para guiar la correcta interpretación del RGPD, sino que son en sí mismos fuente vinculante de obligaciones, contando con fuerza normativa propia. Prueba de ello es que el artículo 33 establece que el artículo 83.5 letra a) establece como una infracción que debe ser sancionada con una administrativa la infracción a "los principios básicos para el tratamiento, incluidas las condiciones para el consentimiento a tenor de los artículos 5, 6, 7 y 9". La utilización del vocablo "incluidas", da cuenta que la mera infracción de los principios básicos para el tratamiento bastaría para constituir una infracción susceptible de ser sancionada con una multa administrativa. No se trata de meros principios orientativos, sino que de verdaderas reglas.

IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

Se trata, entonces, de un catálogo bien definido de principios que orientan a la legislación, pero que también cuentan con autonomía respecto del cumplimiento de sus mandatos por lo que son de aplicación directa, sujetas a sanción por su falta de observancia o incumplimiento (Puente, 2019, p. 117).

## 3.2. Análisis de los principios en la nueva ley

### 3.2.1. Principio de licitud, lealtad y transparencia

El artículo 5.1, letra a) de la nueva ley establece que "Los datos personales serán [...] tratados de manera lícita, leal y transparente en relación con el interesado". Esto implica un triple deber: i) cualquier tratamiento de datos debe realizarse cumpliendo con alguna base de licitud, ii) el tratamiento debe realizarse de buena fe, es decir, de forma leal y sin recurrir a fraude o engaño, y iii) le corresponde al responsable de acreditar la licitud del tratamiento de datos que efectúa (principio de responsabilidad proactiva).

De estos requisitos, es el primero el que más preocupa al análisis de este trabajo, puesto que significa que el requisito de licitud exige que todo tratamiento de datos personales debe sustentarse en al menos una de las bases de legitimidad establecidas en el artículo 6 del RGPD. Cualquier tratamiento que no esté amparado en alguna de esas bases de licitud pasa a ser, por tanto, ilegítimo y contrario a las exigencias del RGPD. Así lo ha señalado la literatura especializada, al recalcar que el principio de licitud "pone especial énfasis en el cumplimiento del ordenamiento jurídico por el responsable del tratamiento, en especial de las bases jurídicas de legitimación de los tratamientos" (Troncoso, 2021, p. 855).



Sin embargo, al revisar las bases de licitud contenidas en el artículo 6 es posible constatar que no resulta del todo claro cuál de ellas es la que resultaría adecuada para fundamentar el tratamiento de datos en el contexto del web scraping para el entrenamiento de modelos de IA.

Una opción sería fundamentar el tratamiento en la causal de la letra a) del artículo 6, es decir, en el hecho que el interesado entregó su consentimiento para el tratamiento de sus datos personales para uno o varios fines específicos. Sin embargo, este mecanismo choca con una barrera operativa: implicaría altísimos costos de transacción para el administrador, quien tendría que obtenerlo uno a uno de cada interesado (Mészáros y Ho, 2018, p. 213).Por otro lado, las condiciones en las que se realiza la actividad de web scraping y el funcionamiento de sistemas de inteligencia artificial , caracterizadas por altos niveles de asimetría de información y falta de transferencia[15], hacen que se vuelva difícil demostrar que se alcanzaron todos los requisitos contenidos en el artículo 32 del GDPR para demostrar que el consentimiento ha sido otorgada de forma válidos, a saber, que este "debe darse mediante un acto afirmativo claro que refleje una manifestación

15. De forma más general, algunos autores han argumentado que distintas prácticas en el mundo digital han erosionado el rol del consentimiento como mecanismo para proteger al interesado. Ver, Andreotta, *et al.* (2021).

IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

de voluntad libre, específica, informada, e inequívoca del interesado". Del mismo modo, el artículo 7 del RGPD establece que el interesado tiene el derecho a retirar su consentimiento en cualquier momento, sin embargo, este derecho se vuelve ilusorio si el interesado no se ha enterado en primer lugar que sus datos están siendo tratados para la finalidad de entrenar modelos de IA.

El ICO del Reino Unido también se ha mostrado escéptico del consentimiento como base de licitud para el tratamiento de datos a través del web scraping, puesto que las organizaciones que entrenan sistemas de IA generativa no cuentan con una relación directa con ellas personas cuyos datos están siendo recolectados. Adicionalmente, el hecho de que los interesados hayan consentido en el uso de sus datos para determinado servicio online de acceso público no implica que hayan podido anticipar que otra empresa pueda luego utilizarlos para fines distintos. Por último, el ICO también manifiesta su preocupación por lo difícil que resulta asegurar a los interesados la capacidad de retirar su consentimiento, puesto que resulta altamente costoso y demandante en términos de tiempo retirar su consentimiento si ello implica que el modelo deberá volver a ser entrenado, una situación que sería altamente costosa (ICO, 2025b, p. 10).

Del mismo modo, el interés legítimo como base de licitud resulta ser una causal de carácter abstracta y que otorga poca certeza jurídica, la cual, si bien se ha discutido a nivel europeo como alternativa para legitimar el procesamiento de datos para entrenar sistemas de inteligencia artificial, todavía está sujeto a interpretaciones divergentes por parte de distintas Autoridades de Protección de Datos (Trigo, 2023). El Supervisor Europeo de Protección de Datos (en adelante "SEPD") señaló que el interés legítimo como base de licitud genera un efecto de imprevisibilidad para los interesados, así como la inseguridad jurídica para los responsables del tratamiento (SEPD, 2024). Por otro lado, la Agencia de Protección de Datos de los Países Bajos ha publicado una guía expresando que las organizaciones e individuos que deseen realizar web scraping primero deben examinar si pueden invocar exitosamente una de las bases de licitud contenidas en la legislación. Generalmente, para ellos, el "interés legítimo" es la única base de licitud a la que podrían optar. Sin embargo, también aclaran la posición de dicha agencia es que el interés legítimo no sería una base de licitud válida para realizar web scraping en aquellos casos en donde el interés del tratamiento sea de carácter exclusivamente comercial (Eutoriteir Persoonsgegevens, 2024).



El 04 de octubre de 2024 el Tribunal de Justicia de la Unión Europea tuvo la oportunidad de referirse a este tema, aclarando que sí es posible invocar la existencia de un interés legítimo para legitimar el tratamiento de datos personales con fines puramente comerciales. Sin embargo, la Corte estableció importantes limitaciones y condiciones para invocar el interés legítimo en estas circunstancias, las que son particularmente relevantes para determinar si el interés legítimo puede ser invocado para justificar la recolección de datos personales a través del web scraping para el entrenamiento de modelos de IA. La Corte señaló que un interés legítimo no necesariamente debe estar establecido por ley, sino que sólo debe ser de carácter lícito, pero además estableció como condición que i) el análisis del test de tres pasos debe realizarse caso a caso ii) dicha evaluación debe incluir el hecho si el interesado puede esperar razonablemente, en el momento y en el contexto de la recogida de los datos personales, que el tratamiento para ese fin pueda tener lugar y iii) el tratamiento debe resultar necesario para

IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

los fines del interés legítimo, por lo que a la luz de las circunstancias particulares los intereses, derechos y libertades de los interesados no prevalezcan sobre el interés legítimo invocado[16].

En particular, el requisito de demostrar que el interesado podía esperar razonablemente la existencia del tratamiento de datos dado el momento y el contexto de la recogida de datos es, a mi parecer, el requisito más difícil de demostrar por parte de los desarrolladores de modelos de IA. En efecto, mucha de la información que se encuentra disponible públicamente en internet se encuentra disponible antes de la emergencia de la tecnología de IA, por lo que difícilmente los interesados se podrían haber representado que su información personas sería recolectada para entrenar dichos sistemas.

Adicionalmente, resulta difícil argumentar que la práctica del web scraping se encuentra tan extendida que cualquier persona debería esperar que cualquier información personal relacionada con su persona que se encuentre disponible en un sitio web es susceptible de ser recolectada para fines indeterminados. Como ha señalado el ICO, la existencia de una práctica común no implica alcanzar las expectativas razonables del interesado y las organizaciones no deberían asumir que el interesado se ha representado el tratamiento solo por tratarse de una práctica extendida en la industria. El ICO es incluso más específico, señalando que lo anterior:

> Aplica particularmente cuando se trata de formas novedosas de utilizar información personal para entrenar modelos de IA generativa a través de mecanismos invisibles o años después de que el interesado proveyó la información o para fines distintos, cuando sus expectativas eran, por defecto, distintas (ICO, 2025b, p.6).



Algunos académicos han criticado el interés legítimo como un retroceso en términos de protección a los individuos y una alternativa al requisito de contar con el consentimiento del interesado que puede afectar el ejercicio de su autodeterminación informativa (Ferretti, 2014). Sin embargo, otros han defendido su aplicación en el contexto del entrenamiento de modelos de inteligencia artificial producto de su flexibilidad, al mismo tiempo que establece resguardos para proteger los derechos de los interesados. Así, Pablo Trigo ha argumentado que este podría invocarse por los desarrolladores de sistemas de IA, en la medida en que se implementen salvaguardas como medidas de seguridad para proteger la información la utilización de anonimización y otras formas de tecnología que promuevan la privacidad (Trigo, 2023, p. 10). En contraposición, Drouard *et al.* identifica importantes desafíos prácticos en el intento de invocar el interés legítimo como base de licitud, en particular la dificultad de implementar medidas efectivas de pseudoanonimización (Druard *et al.*, 2024).

Sin embargo, nuevamente concuerdo con Li, *et al.* en el sentido que la complejidad de la aplicación del interés legítimo y la falta de una guía operacional detallada hace que su aplicación esté plagada de incertidumbre (Li, *et al.*, 2025, p. 5).

A finales de 2024 el Comité Europeo de Protección de Datos (en adelante "CEPD") publicó una guía para abordar este y otros desafíos relacionados a la protección de datos y el desarrollo de sistemas de inteligencia artificial. Su Opinión 28/2024 sobre aspectos

16. Tribunal de Justicia de la Unión Europea. Caso C-621/22 Koninklijke Nederlandse Lawn Tennisbond. ECLI:EU:C:2024:857, considerandos 3, 45 y 55

IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

relacionados con la protección de datos en el contexto de los modelos de IA admite la posibilidad de que el interés legítimo sea invocado como base de licitud en el contexto del desarrollo y puesta a disposición de modelos de IA. Sin embargo, no ofrece una guía detallada que permita a los desarrolladores determinar si, en su caso particular, el web scraping podrá ser amparado por dicha base de licitud.

Adicionalmente, el CEPD repite los requisitos señalados por el Tribunal de Justicia de la Unión Europea en el caso C-621/22. En particular que la necesidad de analizar si el interesado puede esperar razonablemente, en el momento y en el contexto de la recopilación de los datos personales, que se lleve a cabo el tratamiento para dicho fin. El CEPD advierte que los intereses y derechos fundamentales del interesado podrían prevalecer sobre el interés del responsable cuando los datos personales se traten en circunstancias en las que los interesados no esperen razonablemente un tratamiento posterior (CEPD, 2024, punto 47).

Asimismo, advierte que las expectativas razonables de los interesados pueden diferir en función del contexto de la recolección, señalando que esta expectativa será distinta si el desarrollador obtiene los datos a partir de una relación directa con el interesado o si los obtiene de obra fuente, tales como terceras partes o utilizando scraping, dando a entender que en este último caso será más difícil para el desarrollador demostrar el cumplimiento de este requisito (CEPD, 2024, punto 92).

El ICO del Reino Unido, por su parte, ha sufrido un cambio en su posición sobre el interés legítimo a propósito de su proceso de consulta pública. En su publicación de marzo de 2024, llamando a la consulta pública, el ICO aseveró que el interés legítimo podía constituir una base de licitud válida en la medida que los desarrolladores pudiesen demostrar que cumplen con el test de los tres pasos y que implementan salvaguardas técnicas adecuadas (ICO, 2024a). Sin embargo, en su publicación del 10 de diciembre del 2024, en la cual responde a los antecedentes ingresados por los participantes de la consulta pública, la misma entidad se muestra más cautelosa. Así, el ICO explícitamente señaló haber actualizado su posición, dando cuenta que:



> El web scraping es una actividad de procesamiento a gran escala que a menudo ocurre sin que las personas sean conscientes de ello. Este tipo de procesamiento invisible plantea riesgos particulares para los derechos y libertades de las personas. Por ejemplo, si alguien desconoce que sus datos han sido procesados, no puede ejercer sus derechos de información. Hemos recibido evidencia mínima sobre la disponibilidad de medidas de mitigación para abordar este riesgo. Esto significa que, en la práctica, los desarrolladores de IA generativa pueden tener dificultades para demostrar cómo su procesamiento cumple con los requisitos de la prueba de equilibrio de intereses legítimos (ICO, 2024b, p. 3).

La calificación del web scraping como una actividad de alto riesgo constituye una importante carga probatoria para los desarrolladores de sistemas de IA, los que no siempre podrán demostrar que sus intereses prevalecen por sobre los del interesado y que las medidas de mitigación son lo suficientemente robustas.

Todavía no existe jurisprudencia a nivel europeo ni pronunciamientos de agencias regulatorias sobre casos específicos en donde el web scraping busque fundamentar su licitud en el interés legítimo para el entrenamiento de un sistema de IA específico de carácter comercial. Por tanto, es una materia que todavía se encuentra sujeta a los criterios

IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

futuros del regulador y los tribunales. Sin embargo, el triple test del interés legítimo da a entender de qué dependerá mucho del objetivo buscado por el responsable de datos (Contreras y Trigo, 2019). De esta forma, es posible que las Agencias de datos sean más propensas a permitir el web scraping por parte de organizaciones que busquen el interés público, como el Internet Archive, o sean parte esencial del ecosistema de internet, como los motores de búsqueda, en contraposición a empresas que desarrollen sistemas de inteligencia artificial para fines particulares. En especial, teniendo en consideración que pueden existir formas menos lesivas de obtener bases de datos para el entrenamiento de estos sistemas (aunque no necesariamente resultarán comercialmente rentables).

### 3.2.2. Principio de limitación de la finalidad

El artículo 5.1 letra b) define el principio de finalidad limitación de la finalidad como el deber de recoger los datos "con fines determinados, explícitos y legítimos, y no serán tratados ulteriormente de manera incompatible con dichos fines", añadiendo que "el tratamiento ulterior de los datos personales con fines de archivo en interés público, fines de investigación científica e histórica o fines estadísticos no se considerará incompatible con los fines iniciales".

El principio en cuestión se compone de dos elementos. En primer lugar, se encuentra la determinación de la finalidad, y en segundo lugar, la compatibilidad en el uso de los datos. En cuanto al primer aspecto, la finalidad que legitima el tratamiento de los datos debe ser específica, explícita y lícita. El segundo aspecto se refiere a la compatibilidad de uso, lo cual implica que no se podrán realizar tratamientos posteriores de los datos que resulten incompatibles con el propósito originalmente establecido (Troncoso, 2021, p. 859).

Es importante notar que de la redacción del artículo es posible concluir que el principio de limitación de la finalidad se aplica también al "tratamiento ulterior" por parte de todos los terceros a quienes se hayan transferido los datos personales, y no se limita exclusivamente al responsable que los recopiló en primer lugar.

Por tanto, el objetivo de este principio es impedir prevenir la denominada "expansión de la misión" o "function creep" (Koops, 2021), la cual podría resultar en el uso de datos personales más allá de los propósitos originales para los cuales fueron recogidos. Sin embargo, esta limitación no es absoluta. Los datos previamente recopilados pueden ser valiosos para otros fines no especificados inicialmente. Por ello, el principio permite, bajo ciertos parámetros cuidadosamente definidos, un grado limitado de uso adicional mediante una "evaluación de compatibilidad", la cual exige que, en cada caso donde se considere un tratamiento ulterior, se distinga entre aquellos usos adicionales que son "compatibles" y aquellos que deben seguir siendo considerados "incompatibles".

De la lectura del principio no es difícil darse cuenta por qué este se encuentra en tensión o contraposición con la práctica del web scraping. Solove y Hartzog comenta que:

> En clara contradicción con el principio de finalidad, el scraping implica la recopilación indiscriminada de datos con fines no especificados. La mayoría de los fines de los datos recolectados son usos secundarios no relacionados con la finalidad original para la cual los datos fueron recolectados (Solove y Hartzog, 2024, p.34).



IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

A esto habría que añadir que, si bien la legislación permite un uso para una finalidad distinta de la original, esta finalidad debe ser compatible con los autorizados originalmente. Sin embargo, dicho análisis debe ser realizado caso a caso por el responsable de datos que busca utilizar el dato para una finalidad distinta, dicho análisis no será posible en el caso de una recopilación masiva e indiscriminada de grandes volúmenes de datos. En otras palabras, puede que existan casos en que el entrenamiento de sistemas de inteligencia artificial sea una finalidad compatible con la autorizada por el interesado originalmente para la recolección de sus datos personales, pero lo más probable es que dicha hipótesis sea una proporción muy minoritaria de los casos en la práctica.

El ICO aborda específicamente este punto, añadiendo un nivel adicional de complejidad a la evaluación de compatibilidad. De acuerdo con esta entidad, desarrollar un modelo de IA generativo y desarrollar una aplicación específica basada en ese modelo constituyen finalidades distintas. Por tanto, el desarrollo de la aplicación específica basada en el modelo entrenado constituye una finalidad adicional que también debe superar la prueba de compatibilidad. Por tanto, es teóricamente posible que entrenar el modelo de IA generativa cumpla con la prueba de compatibilidad de la finalidad, pero luego desarrollar la aplicación específica basada en dicho modelo no sea compatible con la finalidad que la organización buscó originalmente al momento de recolectar la información a través de web scraping (ICO, 2024b, p. 14).

### 3.2.3. Principio de minimización de datos



El principio de minimización de datos se encuentra definido en el artículo 5.1 letra c) del RGPD, el que prescribe que ""los datos personales serán [...] adecuados, pertinentes y limitados a lo necesario en relación con los fines para los que son tratados". Visto de otra forma, se busca evitar que la recolección de datos resulte excesiva o sobreabundante para los fines buscados y que no se recolecte más información que la realmente necesaria para el objetivo buscado. El considerando 39 del RGPD complementa esta disposición, señalando que "los datos personales sólo deben tratarse si la finalidad del tratamiento no pudiera lograrse razonablemente por otros medios".

Este requisito de necesidad ha sido descrito por la Comisión de Protección de Datos de Irlanda como un requisito que exige demostrar que:

> El procesamiento debe ser un método razonable y proporcionado para lograr un objetivo determinado, teniendo en cuenta el principio general de minimización de datos, y que los datos personales no deben procesarse cuando exista una forma más razonable y proporcionada, y menos intrusiva, de lograr un objetivo (Coimisiún um Chosaint Sonraí, 2019, p. 6).

Así, en el caso Schecke el Tribunal de Justicia de la Unión Europea[17] sostuvo que, al examinar la necesidad del tratamiento de datos personales, el responsable del tratamiento

17. Tribunal de Justicia de la Unión Europea (2010). Casos C-92/09 y C-93/09–Volker und Markus Schecke y Eifert, para 86.

IUS ET SCIENTIA • 2025
Vol. 11 • N° 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

debía considerar medidas alternativas menos intrusivas, y que cualquier injerencia en los derechos de protección de datos derivada del tratamiento en cuestión debía ser la menos restrictiva de dichos derechos. Por tanto, para cumplir con el criterio de necesidad, no debería existir ninguna alternativa igualmente efectiva.

De esta forma se busca prevenir la recolección excesiva de datos personales, asegurando que el uso de dicha información esté debidamente justificado y sea adecuado en relación con el objetivo perseguido. A través de esta norma, se busca proteger los derechos y libertades de los interesados de los datos, reduciendo los riesgos de abuso o gestión indebida de su información personal, ya que no se puede procesar una cantidad excesiva de datos personales que no esté justificada en términos de adecuación, pertinencia y necesidad. Por ejemplo, sería desproporcionado solicitar a un empleado un formulario de salud detallado para verificar su idoneidad para un puesto administrativo que no requiere ningún tipo de aptitud física.

Este es probablemente el principio que se encuentra más en las antípodas con la práctica del web scraping. Mientras el principio de proporcionalidad exige al responsable recolectar el mínimo de información necesaria para alcanzar la finalidad del tratamiento, la práctica del web scraping busca recolectar la mayor cantidad de información posible, para luego encontrar un uso práctico a dicha información. Es por ello que para Solove y Hartzog el principio de proporcionalidad sería la antítesis del web scraping (Solove y Hartzog, 2024, p. 36). Una de las particularidades de los sistemas de inteligencia artificial es que tienen la capacidad de inferir información o correlaciones sorprendentes e inesperadas a partir de conjuntos de datos existentes (Floridi, 2012, p. 1720). Por lo mismo, se vuelve muy difícil para los desarrolladores de sistemas de inteligencia artificial seleccionar un conjunto mínimo de información, puesto que un mayor volumen de datos, en principio, da la posibilidad de encontrar mayores niveles de correlación.



Como contrapunto, un reciente informe del SEPD ha argumentado que el uso excesivo de datos puede incluso perjudicar el desarrollo de sistemas de inteligencia artificial, señalando que:

> El uso de grandes cantidades de datos para entrenar un sistema de IA generativa no implica necesariamente una mayor eficacia o mejores resultados. El diseño cuidadoso de conjuntos de datos bien estructurados, que se utilicen en sistemas que prioricen la calidad sobre la cantidad, siguiendo un proceso de entrenamiento debidamente supervisado y sometido a un seguimiento periódico, es esencial para lograr los resultados esperados, no solo en términos de minimización de datos, sino también cuando se trata de la calidad del resultado y la seguridad de los datos (Supervisor Europeo de Protección de Datos, 2024, p.20).

Si bien esto puede resultar efectivo para el desarrollo de ciertos sistemas de IA, los sistemas como los Large Language Models (LLM) o los modelos de difusión probabilística se benefician de contar con grandes bases de datos que alimenten sus modelos de entrenamiento para la creación de texto e imágenes respectivamente.

Al respecto, es importante diferenciar dos aspectos contenidos en el principio de minimización. El primero es el test de necesidad, es decir, que el procesamiento de datos resulta realmente indispensable para el fin buscado y que este no pueda ser alcanzado por otras vías menos lesivas a los derechos y libertades del interesado. El otro elemento

es que, incluso habiendo demostrado que un procesamiento es indispensable, la forma en que dicho procesamiento se realiza debe demostrar que se recolectan y tratan la menor cantidad posible de datos para alcanzar el fin buscado. Por tanto, podría ser posible que un desarrollador de sistemas de IA demuestre que la única manera posible de entrenar efectivamente su sistema de IA es a través de la recolección de datos a través de web scraping, pero incluso en ese caso deberá demostrar que recolectó la menor cantidad posible datos para el fin buscado. Sin embargo, como hemos señalado anteriormente, muchos sistemas de IA se benefician y vuelven más precisos a mayor cantidad de información que los alimenta, por lo que no es difícil evidenciar la tensión entre la práctica de los desarrolladores de sistemas de IA y los preceptos del principio de minimización de datos.

En cuanto al cumplimiento del test de necesidad es interesante estudiar el cambio de posición que sufrió el ICO del Reino Unido luego de la realización de su ronda de consulta público. En marzo de 2024 subrayó que:

> Actualmente, la mayor parte del entrenamiento de IA generativa solo es posible utilizando el volumen de datos obtenido mediante el scraping a gran escala. Si bien los futuros desarrollos tecnológicos pueden ofrecer soluciones y alternativas novedosas, actualmente hay poca evidencia de que la IA generativa pueda desarrollarse con bases de datos más pequeñas y propietarias (ICO, 2024a).

Sin embargo, en diciembre de 2024 actualizó esta postura, señalando que "Si las organizaciones pueden lograr razonablemente su propósito sin el procesamiento invisible y de alto riesgo que implica el raspado web, entonces no pasarían la parte de necesidad de la prueba de interés legítimo" (ICO, 2024b).



### 3.2.4. Principio de exactitud

El artículo 5.1 letra d) del RGPC establece que los "los datos serán [...] exactos y, si fuera necesario, actualizados; se adoptarán todas las medidas razonables para que se supriman o rectifiquen sin dilación los datos personales que sean inexactos con respecto a los fines para los que se tratan".

Para cumplir este principio los datos deben representar fielmente la situación actual de la persona. Por lo mismo, este principio exige que las organizaciones implementen medidas razonables para garantizar que la información sea adecuada y pertinente respecto del propósito específico para el cual fue obtenida. Sin embargo, esta obligación es difícil de cumplir para aquellos responsables que recolectan datos a través del scraping web, puesto que no mantienen una relación con el interesado, difícilmente podrán actualizar los datos una vez que dejen de ser exactos o pertinentes.

Por otro lado, los mandatos de este principio no se limitan exclusivamente a la calidad de datos, sino que también se extienden al proceso de su tratamiento. En consecuencia, se derivan diversas obligaciones: la actualización constante de la información, la cancelación o eliminación de los datos que no cumplan con los criterios de necesidad o pertinencia en relación con el propósito para el que fueron obtenidos, y la anonimización o minimización de los datos cuando sea requerido. Estas obligaciones buscan asegurar que los datos se mantengan dentro de los parámetros adecuados y

IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

pertinentes a lo largo de su ciclo de vida. Al recolectar un volumen tan alto de información disponible públicamente, a partir de fuentes tan diversas, resulta poco escalable y realista exigir al responsable que realiza web scraping para el entrenamiento de sistemas de inteligencia artificial que cumpla con el principio de calidad en los términos antes expuestos.

Adicionalmente, la exactitud, actualidad y pertinencia de los datos no sólo debe contrastarse con la situación actual del interesado, sino que además tiene relación con la proveniencia y la finalidad del tratamiento de los datos. Como hemos mencionado anteriormente, para el desarrollador que realiza web scraping resulta poco realista y escalable realizar este análisis e incluso tener en consideración cuál fue la finalidad de la recolección de datos que está siendo obtenido de fuentes públicas en internet.

Por último, desde un punto de vista técnico es importante tener en consideración que, incluso en aquellos casos en que los sistemas de inteligencia artificial son entrenados con información de alta calidad, estos son propensos a entregar resultados inexactos y sufrir lo que la literatura ha denominado "alucinaciones". Dichas alucinaciones tienen lugar cuando el sistema entrega una respuesta ficticia, errónea o información no sustentada al responder un requerimiento (Kumar *et al.*, 2023). Esto puede dar pie a que los datos personales recolectados sean asociados a información que no es verídica, esté sacada de contexto o simplemente carezca de sentido.

### 3.2.5. Principio de responsabilidad proactiva



Este principio, contenido en el artículo 5.2 del RGPD establece que "El responsable del tratamiento será responsable del cumplimiento de lo dispuesto en el apartado 1 y capaz de demostrarlo". Por tanto, no resulta particularmente relevante para el análisis de este trabajo, puesto que sólo recalca que los responsables de cumplir con las obligaciones contenidas en el Reglamento y responder por las infracciones cometidas. Sí es relevante en el sentido que dicho principio no podrá ser cumplido por los responsables que realicen web scraping, por el simple hecho que tampoco es posible que cumplan los otros principios establecidos en el artículo 5, en particular el principio de limitación de la finalidad y minimización de datos.

### 3.2.6. Principio de integridad y confidencialidad

El artículo 5.1 letra f) señala que "los datos personales serán [...] tratados de tal manera que se garantice una seguridad adecuada de los datos personales, incluida la protección contra el tratamiento no autorizado o ilícito y contra su pérdida, destrucción o daño accidental, mediante la aplicación de medidas técnicas u organizativas apropiadas"".

Esto genera un interesante debate en torno al web scraping ¿ha faltado a su deber de integridad y confidencialidad el administrador de un sitio web disponible públicamente en internet cuya información es objeto de recolección a través de web scraping? Intuitivamente podemos responder que no, puesto que la información ya se encontraba públicamente disponible en un comienzo.

IUS ET SCIENTIA • 2025
Vol. 11 • N° 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

Sin embargo, una declaración conjunta sobre data scraping y protección de datos personales suscrita por los jefes de varias autoridades en materia de protección de datos personales buscó aclarar los siguientes puntos:

1. La información personal que se encuentra públicamente accesible todavía está sujeta a las obligaciones de privacidad y protección de datos personales.
2. Los administradores de sitios web y redes sociales tienen la obligación de proteger los datos personales en sus plataformas de la recolección ilícita de dicha información a través de data scraping y,
3. La recolección masiva de información a través de data scraping que recolecta información personal puede constituir un incidente de fuga de datos que debe ser reportado,

La declaración resalta que la información recolectada a través de web scraping puede ser explotada para distintos fines, como la monetización a través de la reutilización en sitios web de terceros, la venta a actores maliciosos o el análisis privado o la recopilación de inteligencia, lo que genera graves riesgos para las personas (Office of the Privacy Commissioner of Canada, 2023).

Así, por ejemplo, una Corte Federal de Justicia alemana recientemente sancionó a la red social Facebook por no implementar medidas adecuadas de seguridad, al permitir que el número de teléfono de un individuo fuese públicamente accesible a terceros a través de la función para importar contactos. Terceros ajenos al interesado utilizaron esta función para realizar consultas aleatorias y obtener acceso a información del usuario, como su nombre, género y empleador[18]. En este caso, a pesar de que la información se encontraba públicamente disponible, la Corte llegó a la conclusión que Facebook incumplió sus deberes de seguridad al admitir que la información fuese objeto de web scraping por parte de terceros, producto de la afectación que esto produjo al interesado.



## IV. Reglamento (UE) 2024/1689, sobre inteligencia artificial

Hasta el momento he descrito cómo el requisito de poder demostrar la existencia de una base de licitud para el tratamiento de datos personales, así como el cumplimiento de los principios de protección de datos personales contenidos en el RGPD pone a los desarrolladores de sistemas de inteligencia artificial en una posición de incertidumbre jurídica, en la que les es difícil determinar si están obrando de forma lícita al recolectar información a través del web scraping para el entrenamiento de sus modelos de IA. Ahora corresponde analizar si el RIA contiene alguna solución regulatoria que permita a los desarrolladores seguir una guía clara para determinar cuándo resulta lícito la recolección de datos para el entrenamiento de sistemas de inteligencia artificial.

El Reglamento de IA fue publicado el 13 de junio de 2024. Este cuerpo normativo es uno de los primeros instrumentos a nivel global dedicado a la gobernanza de la inteligencia artificial (Almada, 2025). El cuerpo normativo adopta una aproximación basada

18. BGH Alemania (2024). Facebook/Meta VI ZR 10/24.

IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

en la gestión de riesgos, clasificando los sistemas de inteligencia artificial en sistemas de distintos niveles de riesgo. Estos distintos niveles de riesgo tienen asociados distintos niveles de obligaciones, tales como gestión de riesgo, gobernanza de datos, documentación técnica, conservación de registros, supervisión humana y ciberseguridad (Llano Alonso, 2024). Al mismo tiempo, se listan prácticas de IA prohibidas, las que se encuentran derechamente prohibidas, tanto respecto de introducción en el mercado, la puesta en servicio o su utilización.

A fin de mantener un equilibrio entre la importante carga regulatoria que las organizaciones de la UE deberán asumir en el cumplimiento de los requisitos contenidos en el Reglamento de IA, este también cuenta con una batería de medidas de apoyo a la innovación, contenidas en los artículos 57 a 63 del RIA.

Para efectos de este trabajo nos interesa especialmente la figura de los espacios controlados de pruebas para la IA, las que son definidas por el artículo 3 N° 55 como

> un marco controlado establecido por una autoridad competente que ofrece a los proveedores y proveedores potenciales de sistemas de IA la posibilidad de desarrollar, entrenar, validar y probar, en condiciones reales cuando proceda, un sistema de IA innovador, con arreglo a un plan del espacio controlado de pruebas y durante un tiempo limitado, bajo supervisión regulatoria.

El objetivo de estos espacios es proporcionar "un entorno controlado que fomente la innovación y facilite el desarrollo, el entrenamiento, la prueba y la validación de sistemas innovadores de IA" (artículo 57.5).



Participar de estos espacios controlados puede entregar distintos beneficios a los desarrolladores de sistemas de IA de alto riesgo. Entre ellos, el hecho que las autoridades competentes están obligadas a orientar, supervisar y apoyar al desarrollador dentro del espacio controlado de pruebas para la IA a fin de determinar los riesgos y las medidas de reducción en relación con las obligaciones del Reglamento (artículo 57.6). Del mismo modo, las autoridades estarán obligadas a orientar a los participantes del espacio controlado de pruebas "sobre las expectativas en materia de regulación y la manera de cumplir los requisitos y obligaciones establecidos en el presente Reglamento".

Del mismo modo, el RIA fomenta la participación de las autoridades nacionales competentes en materia de protección de datos personales en estos espacios controlados de prueba para la IA, en la medida en que lo permitan sus respectivas funciones y competencia (artículo 57.10). En este sentido, los desarrolladores que participen de estos espacios tendrán la ventaja de obtener asesoramiento personalizado sobre el cumplimiento de las obligaciones regulatorias directamente de las mismas autoridades supervisoras que se encargan de hacer cumplir las reglas pertinentes (Baldini y Francis, 2024, p. 7).

Sin duda el mayor atractivo de estos espacios de prueba es que los desarrolladores, cumpliendo ciertas condiciones, se encontrarán exentos de la imposición de multas administrativas por el incumplimiento de las disposiciones del RIA. Así, el artículo 57. 12 señala que

> siempre que los proveedores potenciales respeten el plan específico y las condiciones de su participación y sigan de buena fe las orientaciones proporcionadas por la

IUS ET SCIENTIA • 2025
Vol. 11 • N° 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

> autoridad nacional competente, las autoridades no impondrán multas administrativas por infracciones del presente Reglamento. En los casos en que otras autoridades competentes responsables de otras disposiciones del Derecho de la Unión y nacional hayan participado activamente en la supervisión del sistema de IA en el espacio controlado de pruebas y hayan proporcionado orientaciones para el cumplimiento, **no se impondrán multas administrativas en relación con dichas disposiciones**[19].

Esto abre la puerta para que aquellos desarrolladores de sistemas de inteligencia artificial que quieran entrenar sus modelos de IA a partir de información recolectada a través de web scraping lo realicen en el contexto de un espacio controlado de pruebas. Esto le permitiría obtener guía específica y detallada respecto del cumplimiento normativo por parte de las distintas autoridades involucradas e incluso encontrarse exento de la aplicación de multas administrativas.

Sin embargo, en principio los espacios controlados de prueba están diseñados para hacer más laxos los requisitos regulatorios del RIA en el período previo a la introducción del sistema en el mercado. Por lo mismo, en principio los desarrolladores que se encuentren participando de estos espacios deberán cumplir a cabalidad con todas las exigencias regulatorias establecidas en el RGPD. Si bien es posible que la autoridad nacional o local de control en materia de protección de datos partícipe del espacio controlado de prueba, de acuerdo con lo establecido en el artículo 57.10, y por tanto exima al desarrollador de la interposición de multas administrativas por aplicación del RGPD, esta circunstancia no necesariamente se configurará. En primer lugar, porque el artículo 57.10 utiliza un lenguaje que deja claro que la participación de las autoridades de protección de datos personales en el espacio controlado de pruebas no es de carácter obligatorio, sino que es una circunstancia por la cual se deberá "velar". Segundo, porque dicha participación está condicionada a que así lo permitan sus respectivas funciones y competencias de la autoridad de control de protección de datos personales.



Más relevante aún es la limitación temporal aplicada a la exención de la imposición de infracciones en los espacios controlados de prueba, puesto que estas sólo son aplicables durante un período limitado antes de su introducción en el mercado o su puesta en servicio, con arreglo a un plan del espacio controlado de pruebas específico acordado entre los proveedores o proveedores potenciales y la autoridad competente (57.5). Por tanto, incluso en aquellos casos en que durante la etapa de entrenamiento el modelo de IA haya estado exento de las obligaciones contenidas en el RGPD respecto a contar con una base de licitud y cumplir con los principios de la regulación de datos personales, dichos requisitos sí serán plenamente aplicables una vez que dicho sistema sea desplegado y puesto a disposición en el mercado. Esto es particularmente complejo, puesto que muchos sistemas de IA requieren constantemente ser reentrenados con nueva información para su aprendizaje reforzado.

19. El énfasis es mío.

IUS ET SCIENTIA • 2025
Vol. 11 • N° 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

## V. Alternativas al cumplimiento de los principios de la protección de datos personales

Prohibir a los desarrolladores de sistemas de IA realizar web scraping implicaría un importante costo social y económico, así como un obstáculo importante en el avance del campo de la inteligencia artificial. Así, Google ha declarado que de prosperar una de las demandas que buscan prohibir el web scraping en Estados Unidos, esto no sólo significaría un mazazo para los servicios de Google, sino que para el desarrollo de la IA generativa en general (Reuters, 2023).

Por tanto, es necesario alcanzar una solución regulatoria capaz de compatibilizar la capacidad de los desarrolladores para obtener datos que les permitan entrenar a sus sistemas de inteligencia artificial con la debida protección de la autodeterminación informativa de los interesados de datos personales.

Una alternativa que merece ser estudiada con mayor detención en un trabajo futuro es diseñar un sistema de permisos administrativos que permita a las autoridades de protección de datos personales otorgar una licencia o permiso administrativo para la realización de web scraping por parte de un desarrollador de sistemas de inteligencia artificial, luego de haber realizado una evaluación del caso concreto y decretado las medidas de salvaguarda y compensación correspondientes. Bajo este modelo se podría compatibilizar la necesidad de los sistemas de inteligencia artificial de realizar web scraping para entrenar sus algoritmos y a la vez cautelar los derechos de los interesados de datos personales. Adicionalmente, esta autorización administrativa puede permitir a los desarrolladores de sistemas de inteligencia artificial eludir legítimamente las medidas de exclusión que muchos sitios web están implementando para evitar ser objetos de web scraping, descritas en la sección 2.5 de este trabajo, ya sea en sus términos y condiciones o en la capa técnica a través del uso de Robots Exclusion Protocol (REP). Esto permitiría entregar una ventaja tangible a las organizaciones que decidan optar por este permiso administrativo, otorgándole un acceso más amplio a datos de entrenamiento al mismo tiempo que se le entregan mejores condiciones de certeza jurídica. Esta autorización deberá estar acompañada de importantes compromisos de confidencialidad, seguridad y otras medidas de mitigación.

Más importante aún, permitiría que el criterio de autorización no esté basado en hipótesis abstractas contenidas en la ley, sino que en un procedimiento administrativo que busque optimizar el interés público y el control democrático por sobre el desarrollo de sistemas de inteligencia artificial. Hasta el momento la protección de datos personales ha estado enfocada en la protección de los interesados como individuos, sin embargo, el desarrollo de la inteligencia artificial da cuenta de la necesidad ir más allá de la protección del interés individual del interesado, y reconoce la existencia de un interés general en la suma de intereses individuales de los interesados.

Esta propuesta de regulación ex-ante busca reconocer al estado como representante y garante del interés general y del ejercicio del control democrático sobre las actividades de los particulares (Chan *et al.*, 2023). De esta forma, se podría potenciar que el desarrollo de sistemas de inteligencia artificial no sólo esté dejado a la deriva de los intereses del mercado, sino que la sociedad en su conjunto tenga injerencia respecto



IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

a qué tipo de inteligencia artificial es desarrollada, tanto desde una perspectiva económica, tecnológica y social, y cuáles son los valores que esta debe reflejar.

Sin embargo, el diseño y contenido de un sistema de estas características excede con creces el objetivo de este trabajo y deberá ser objeto de una futura investigación.

## 6. Conclusiones preliminares

1. El web scraping implica la recolección automatizada de información contenida en fuentes públicas en internet, a través del uso de scripts o bots. Dicha práctica ha cumplido distintas funciones en distintas etapas de la evolución de internet.
2. El entrenamiento de sistemas de inteligencia artificial requiere de enormes volúmenes de datos, los que muchas veces son obtenidos de fuentes públicas en internet a través del web scraping.
3. El GDPR cuenta con estrictos requisitos respecto a la licitud del tratamiento de datos personales, así como el cumplimiento de principios para el tratamiento de datos, en participar el principio de limitación de la finalidad y el principio de minimización de datos.
4. A partir del análisis realizado en este trabajo es posible concluir que la práctica del web scraping es incompatible con la mayoría de los principios de la legislación de protección de datos personales. En particular, el hecho que el web scraping implica la recolección masiva e indiscriminada de datos personales para el entrenamiento de sistemas de inteligencia artificial pone dicha actividad en directa contraposición con los principios de minimización de datos y limitación de la finalidad.
5. Si bien el RIA ha creado medidas para el apoyo a la innovación, como los espacios controlados de prueba, dichos espacios controlados sólo aplican para la etapa de entrenamiento de los sistemas de IA y no necesariamente los eximen del cumplimiento de las disposiciones del RGPD.
6. Corresponde al legislador alcanzar una solución regulatoria que sea capaz de compatibilizar el desarrollo de sistemas de inteligencia artificial con la protección de la autodeterminación informativa de los interesados de datos personales.
7. El trabajo esboza una propuesta para alcanzar dicho equilibrio: a través de un sistema de permisos administrativos que habilite a los desarrolladores a realizar web scraping para obtener datos para entrenar sus modelos de IA, incluso eludiendo medidas técnicas de exclusión implementadas por los sitios web. Sin embargo, dicha discusión escapa el objetivo de este trabajo y deberá ser objeto de una futura investigación.



IUS ET SCIENTIA • 2025
Vol. 11 • N° 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

## Bibliografía

Akhtar, F. (2023). Regulating Artificial Intelligence for a Safer and More Ethical Future: A Review of the EU's AI Act. http://dx.doi.org/10.2139/ssrn.4560224

Andreotta, *et al.* (2021). AI, big data, and the future of consent. *AI & Society*. Volume 37, p. 1715–1728. https://doi.org/10.1007/s00146-021-01262-5

Alemohammad, S. *et al.* (2023). Self-consuming generative models go MAD. ArXiv, abs/2307.01850.

Almada, M, (2025). The EU AI Act in a Global Perspective. Handbook on the Global Governance of AI (Furendal & Lundgren, eds, Edward Elgar 2025), http://dx.doi.org/10.2139/ssrn.5083993

Almaqbali, I. S., Al Khufairi, F. M., Khan, M. S., Bhat, A. Z., Ahmed, I. (2019). Web Scraping: Data Extraction from Websites. *Journal of Student Research*. https://doi.org/10.47611/jsr.vi.942

Baack, S. (2024). A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl', The 2024 ACM Conference on Fairness, Accountability, and Transparency https://dl.acm.org/doi/10.1145/3630106.3659033

Baldini, D. y Francis, K. (2024). AI Regulatory Sandboxes between the AI Act and the GDPR: the role of Data Protection as a Corporate Social Responsibility. Conference: ITASEC 2024 Italian Conference on Cyber Security 2024. https://doi.org/10.2139/ssrn.5533498

Bhatia, M. A. (2016). Artificial Intelligence–Making an Intelligent personal assistant. Indian J. Comput. Sci. Eng, 6, 208-214.

Birhane, *et al.*(2023) Into the LAIONs Den: Investigating Hate in Multimodal Datasets. arXiv, http://arxiv.org/abs/2311.03449

Blake, B. (2023), Google Says Data-Scraping Lawsuit Would Take 'Sledgehammer' to Generative AI [recurso web]. Publicado el 17 de octubre de 2023 en Reuters https://www.reuters.com/legal/litigation/google-says-data-scraping-lawsuit-would-take-sledgehammer-generative-ai-2023-10-17/ (último acceso el 21 de marzo de 2025)

Bowles *et al.*(2018). GAN Augmentation: Augmenting Training Data Using Generative Adversarial Networks, arXiv https://doi.org/10.48550/arXiv.1810.10863

Chan, A., Bradley, H. y Rajkumar, N. (2023). Reclaiming the Digital Commons: A Public Data Trust for Training Data. Accepted at AIES 2023 https://doi.org/10.48550/arXiv.2303.09001

Comité Europeo de Protección de Datos (2024). Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models [recurso web]. Disponible en: https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en (último acceso el 21 de marzo de 2025).

Contreras, P. y Trigo, P. (2019). Interés legítimo y tratamiento de datos personales: Antecedentes comparados y regulación en Chile. *Revista Chilena De Derecho Y Tecnología*, 8(1), 69–106. https://doi.org/10.5354/0719-2584.2019.52915

Cook, S. (2008) The contribution revolution, *Harvard Business Review*, 86, 10, 60-69.

de Terwangne, C. (2020). Article 5. Principles relating to processing of personal data, en Kuner, Christopher *et al.* (eds.), The EU General Data Protection Regulation (GPDR). A Commentary (Oxford, Oxford University Press). https://doi.org/10.1093/oso/9780198826491.003.0034

Drouard, E., *et al.* (2024). The Interplay between the AI Act and the GDPR: Part I – When and How to Comply with Both. *Journal of AI Law and Regulation*, Volume 1, Issue 2, pp. 164-176. https://doi.org/10.21552/aire/2024/2/4



IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

EUneedsAI (2024). AN OPEN LETTER Europe needs regulatory certainty on AI [recurso web]. Disponible en: https://euneedsai.com/?utm_source=substack&utm_medium=email#signatories (último acceso el 21 de marzo de 2025).

Eutoriteir Persoonsgegevens (2024) Scraping door particulieren en private organisaties. Informe disponible en: https://autoriteitpersoonsgegevens.nl/system/files?file=2024-05/Handreiking%20scraping%20door%20particulieren%20en%20private%20organisaties.pdf (último acceso el 21 de marzo de 2025)

Ferretti, F. (2014). Data Protection and the Legitimate Interest of Data Controllers: Much Ado about Nothing or the Winter of Rights? 51 *Common Market Law Review*, Volume 51, Issue 3, pp. 843 – 868. https://doi.org/10.54648/COLA2014063

Financial Times (2024). AI start-up Anthropic accused of 'egregious' data scraping [recurso web]. Disponible en: https://www.ft.com/content/07611b74-3d69-4579-9089-f2fc2af-61baa?ref=platformer.news (último acceso el 21 de marzo de 2025)

Floridi, L. (2012). Big data and their epistemological challenge. Philos Technol 25:435–437. https://doi.org/10.1007/s13347-012-0093-4

Folberth, A. Jahnel, J. Bareis, J. Orwat, C. y Wadephul, C. (2022). Tackling Problems, Harvesting Benefits: A Systematic Review of the Regulatory Debate Around AI. *Karlsruher Institut für Technologie (KIT)*. https://doi.org/10.5445/IR/1000150432.

Guadamuz, A. (2024). A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs, *GRUR International*, Volume 73, Issue 2, pp 111–127, https://doi.org/10.1093/grurint/ikad140

An Coimisiún um Chosaint Sonraí (2019) Guidance Note: Legal Bases for Processing Personal Data [recurso web]. Disponible en: https://www.dataprotection.ie/sites/default/files/uploads/2020-04/Guidance%20on%20Legal%20Bases.pdf (último acceso el 21 de marzo de 2025).

Hacker, P. (2021). A legal framework for AI training data–from first principles to the Artificial Intelligence Act. *Law, Innovation and Technology*, Vol 13, No. 2, 257-301. https://doi.org/10.1080/17579961.2021.1977219

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines,* 30 (1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

HmbBfDI (2024). Discussion Paper: Large Language Models and Personal Data [recurso web]. Disponible en: https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Discussion_Paper_Hamburg_DPA_KI_Models.pdf (último acceso el 21 de marzo de 2025).

Henrys, K (2021). Importance of web scraping in e-commerce. http://dx.doi.org/10.2139/ssrn.3769593

IAPP (2023). Training AI on personal data scraped from the web [recurso web]. Disponible en: https://iapp.org/news/a/training-ai-on-personal-data-scraped-from-the-web (último acceso el 21 de marzo de 2025).

Information Commissioner's Office (2024a). Generative AI first call for evidence: The lawful basis for web scraping to train generative AI models [recurso web]. Disponible en: https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/generative-ai-first-call-for-evidence/ (último acceso el 21 de marzo de 2025).

Information Commissioner's Office (2024b). Information Commissioner's Office response to the consultation series on generative AI [recurso web]. Disponible en:https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/response-to-the-consultation-series-on-generative-ai/ (último acceso el 21 de marzo de 2025).



IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

Jayachandran, J. y Arni, V. (2023). Traversing the Ethical Landscape of Data Scraping for AI http://dx.doi.org/10.2139/ssrn.4666354

Jobin, A. Ienca, M. y Vayena, E. (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1 (9), 389–99.

Khder, M. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing and its Applications*. https://doi.org/10.15849/IJASCA.211128.11

Koops, B. (2021). The concept of function creep. *Law, Innovation and Technology*. 13. 1-28. https://doi.org/10.1080/17579961.2021.1898299

Kumar, M. *et al.* (2023). Artificial Hallucinations by Google Bard: Think Before You Leap. *Cureus*. 15. https://doi.org/10.7759/cureus.43313.

Llano Alonso, F. (2024). Artículo 14. Supervisión Humana, en "Comentarios al Reglamento Europeo de Inteligencia Artificial" Moisés Barrio Andrés (director). Editorial La Ley (Madrid).

Krotov, V y Silva, L., (2018). Legality and ethics of web scraping, Twentyfourth Americas Conference on Information Systems, New Orleans.

Medina, M. (2022). El derecho a conocer los algoritmos utilizados en la toma de decisiones. Aproximación desde la perspectiva del derecho fundamental a la protección de datos personales. *Teoría y realidad constitucional*, ISSN 1139-5583, Nº 49.

Mészáros y Ho (2018). Big Data and Scientific Research. 59 *Hungarian Journal of Legal Studies* 403 (405). https://doi.org/10.1556/2052.2018.59.4.5

Milev, P. (2017). Conceptual approach for development of web scraping applications for tracking information. *Economic Alternatives*, (3), 475-485.

Moerel, L. y Storm, M. (2024). Do LLMs "store" Personal Data? This Is Asking the Wrong Question [recurso web]. Disponible en: https://iapp.org/news/a/do-llms-store-personal-data-this-is-asking-the-wrong-question (último acceso el 21 de marzo de 2025).

Neel, S. (2024). Privacy Issues in Large Language Models: A Survey. arXiv https://doi.org/10.48550/arXiv.2312.06717

Nissenbaum, H (2011). A Contextual Approach to Privacy Online. *Daedalus* 140 (4), Fall 2011: 32-48, https://doi.org/10.1162/DAED_a_00113

Li, W. *et al.* (2025) The Quest for Lawful AI Training under Data Protection Frameworks: Global Controversies and Practical Implication http://dx.doi.org/10.2139/ssrn.5162653

Longpre, S. *et al.* (2024). Consent in Crisis: The Rapid Decline of the AI Data Commons. Cornell University https://doi.org/10.48550/arXiv.2407.14933

OECD (2025). Intellectual Property Issues in Artificial Intelligence Trained of Scraped Data. OCD Artificial Intelligence Papers N° 33. https://doi.org/10.1787/7b245f7e-en

Office of the Privacy Commissioner of Canada (2023). Joint statement on data scraping and the protection of privacy, publicada el 24 de agosto de 2023 [recurso web]. Disponible en: https://www.priv.gc.ca/en/opc-news/speeches-and-statements/2023/js-dc_20230824/ (último acceso el 21 de marzo de 2025).

O'Reilly, T. (2007). What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies*, No. 1, p. 17, First Quarter 2007.

O'Reilly, T. y Battele, J. (2009) Web squared: Web 2.0 five years on, O'Reilly and TechWeb [recurso web]. Disponible en: https://www.kimchristen.com/wp-content/uploads/2015/07/web2009_websquared-whitepaper.pdf (último acceso el 21 de marzo de 2025).

Puente Escobar, A. (2019). Principios y licitud del tratamiento, en Rallo Lombarte, Artemi (dir.), Tratado de Protección de Datos (Valencia, Tirant lo Blanch).



IUS ET SCIENTIA • 2025
Vol. 11 • Nº 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04

Regine, P. (2022). The Politics of Regulating Artificial Intelligence Technologies: A Competition State Perspective. *Handbook on Public Policy and Artificial Intelligence*, edited by Regine Paul, Emma Carmel and Jennifer Cobbe (Cheltenham Spa: Edward Elgar).

Reuters (2023). Google Says Data-Scraping Lawsuit Would Take 'Sledgehammer' to Generative AI, REUTERS [recurso web]. Disponible en: https://www.reuters.com/legal/litigation/google-says-data-scraping-lawsuit-would-take-sledgehammer-generative-ai-2023-10-17/ (último acceso el 10 de noviembre de 2025)

Sellars, A, (2018). Twenty Years of Web Scraping and the Computer Fraud and Abuse Act. *Scholarly Commons* at Boston University School of Law.

Shumailov, I. *et al.* (2024) AI models collapse when trained on recursively generated data. *Nature* 631, 755–759 https://doi.org/10.1038/s41586-024-07566-y

Sirisuriya, D. S., (2015). A comparative study on web scraping. Proceedings of 8th International Research Conference, KDU.

Solove, D. y Hartzog, W. (2024). The Great Scrape: The Clash Between Scraping and Privacy. 113 *California Law Review* 1521. https://doi.org/10.2139/ssrn.4884485

Supervisor Europeo de Protección de Datos (2024). La IA generativa y el EUDPR: Primeras orientaciones del SEPD para garantizar el cumplimiento de la protección de datos al utilizar sistemas de IA. Disponible en: https://www.edps.europa.eu/system/files/2024-06/24-06-03_genai_orientations_en.pdf (último acceso el 21 de marzo de 2025).

Tribunal de Justicia de la Unión Europea (2010). Casos C-92/09 y C-93/09–Volker und Markus Schecke y Eifert.

Tribunal de Justicia de la Unión Europea. Caso C-621/22 Koninklijke Nederlandse Lawn Tennisbond. ECLI:EU:C:2024:857

Trigo, P. (2023). Can legitimate interest be an appropriate lawful basis for processing Artificial Intelligence training datasets? *Computer Law & Security Review* 48–105765. https://doi.org/10.1016/j.clsr.2022.105765

Troncoso, A. (2021). Los principios relativos al tratamiento (comentario al artículo 5 RGPD y al artículo 4 LOPDGDD), en Troncoso Reigada, Antonio (dir), Comentario al Reglamento General de Protección de Datos ya la Ley Orgánica de Protección de Datos Personales y Garantía de los Derechos Digitales (Madrid, Civitas–Thomson Reuters).

Tschider, C. (2021) AI's Legitimate Interest: Towards a Public Benefit Privacy Model, 21 *Hous. J. Health L. & Policy* 125, 132

Viollier, P. (2021). Taming the Algorithm: Analyzing EU Enforcement Mechanisms to Enhance Algorithmic Transparency and Accountability [Tesis de maestría, Leiden Law School].

Wilson, D. Lin, X. Longstreet, P. y Sarker, S. (2011). Web 2.0: A Definition, Literature Review, and Directions for Future Research. AMCIS 2011 Proceedings–All Submissions. 368. http://aisel.aisnet.org/amcis2011_submissions/368

Yang Sun, Z, y Lee Giles, C. (2007). A large-scale study of robots.txt. In Proceedings of the 16th international conference on World Wide Web (WWW '07). Association for Computing Machinery, New York, NY, USA, 1123–1124. https://doi.org/10.1145/1242572.1242726



IUS ET SCIENTIA • 2025
Vol. 11 • N° 2 • pp. 74-103
ISSN 2444-8478 • https://dx.doi.org/10.12795/IETSCIENTIA.2025.i02.04