LaNbarque
Abogados
CENTRADOS EN TUS INTERESES

ALEGACIONES A LAS DIRECTRICES 03/2026, DEL COMITÉ EUROPEO DE PROTECCIÓN DE DATOS, SOBRE *WEB SCRAPING* EN EL CONTEXTO DE LA INTELIGENCIA ARTIFICIAL GENERATIVA

La claudicación del derecho fundamental a la protección de datos frente a la dialéctica del mercado: de la ponderación al abatimiento del derecho

1. LA DERROTA DEL DERECHO FUNDAMENTAL A LA PROTECCIÓN DE DATOS EN EL MARCO DEL *BIG DATA.*

La técnica *web scraping* consiste en la extracción o recopilación automatizada de ingentes cantidades de datos disponibles en los servicios web para transformar estos datos no estructurados en datos estructurados para ser tratados para fines diversos.

Esta técnica que ha adquirido una relevancia tremendamente significativa en el contexto del entrenamiento de los modelos de la IA generativa. Sin embargo, su utilización arrancó con el inicio de la eclosión de Internet en los años 90, cuando los motores de búsqueda comenzaron a indexar sitios webs. Sin embargo, ha de efectuar una distinción esencial, una cosa es que los motores de búsqueda efectúen rastreo de la web, lo cual, a priori, puede ser beneficioso para los motores de búsqueda y sus usuarios que hacen que la información sea accesible en el espacio web -y relativamente sencillo ejercer sus derechos como el de oposición u olvido- y otra es que otras empresas realicen el almacenamiento de datos, sin consentimiento y sin transparencia para fines ni siquiera conocidos por los titulares de los datos[1]. No obstante, si la privacidad y la actividad de los motores de búsqueda ya plantean desafíos para la privacidad, el problema se agrava definitivamente con el scraping web y el entrenamiento de los modelos de la IA.

[1] Solove J., Daniel y Woodrow Hartzog, "*The Great Scrape: The Clash Between Scraping and Privacy*", California Law Review, Vol.113, p.1521

Se trata, en consecuencia, de una actividad de tratamiento masivo, de carácter global y sin conocimiento de los interesados, lo cual supone un riesgo adicional y esencial para los derechos y libertades. Ya en 2017, el informe Comité Internacional de Bioética de la UNESCO, "Big Data y Salud" destacó lo que:

> *"La escala y la complejidad del procesamiento de datos personales en el área del Big Data hace prácticamente imposible que las personas realicen un seguimiento o tomen decisiones significativas sobre todos los usos de sus datos personales… Esto conduce a una pérdida de autonomía que, en consecuencia, significa una pérdida de control y a una pérdida de libertad frente a las decisiones tomadas por el entorno tecnológico y los procesos automatizados. La gobernanza tiene que compensar esta pérdida de control, garantizando que todas las dimensiones de la autonomía sean respetadas y apoyadas"*[2].

La primera consideración crítica que cabe realizar es, precisamente, que por deferencia se acomete una redacción con carencia de contexto. No solamente desde el punto de vista tecnológico, donde la evolución de la IA y el *scraping* supone un nuevo reto para la protección de datos, sino por la falta identificación del calado jurídico y ético que conlleva esta técnica de tratamiento de datos de carácter personal.

Así, por ejemplo, las grandes plataformas de redes sociales son recursos tremendamente estimables para el rastreo de datos donde se insertan datos de la vida cotidiana tales como conversaciones, preferencias, opiniones, amistades, fotos, vídeos, ubicación, etc. El entrenamiento de los modelos de IA generativa precisa de un entrenamiento sin pausa con volúmenes inconmensurable de datos en la que los operadores han entrado en una competición por su acaparación.

Debería existir una expectativa razonable de privacidad de manera que nuestras fotos no estuvieran al servicio de sistemas de identificación facial, los datos de nuestro estilo de vida no fueran utilizados para realizar cálculos actuariales de nuestra prima del seguro, nuestras

[2]Párrafo 48. SHS/SÍ/IBC-24/17/3 REV.2. París, 15 de septiembre de 2017

opiniones no sirvieran para evaluar nuestra empleabilidad, etc. O, modular la propaganda política de acuerdo con nuestro perfilado previo a fin de realizar campañas personalizadas de desinformación, lo cual supone un riesgo para el funcionamiento democrático de la sociedad.

En el marco del *big data* en el momento actual de los prolegómenos de la aceleración de la inteligencia artificial los riesgos de esta técnica para la efectividad del derecho fundamental de protección de datos son tan elevados que puede suponer el desmoronamiento del contenido esencial del derecho.

La configuración por el derecho de la Unión del derecho a la protección de datos de carácter personal presenta *ad initio* un importante sesgo de diseño en favor de la lógica de mercado. Con la intensificación del desarrollo digital este exceso de ponderación se ha acentuado a fin de viabilizar la economía digital.

Una interpretación normativa deferente es incapaz de sujetar al sector a la regla del derecho. Todo ello en el marco del oligopolio global de las *Big Tech* conformado por un reducido grupo de corporaciones de Estados Unidos y de China que se presentan como poderes privados inéditos.

La carencia de contexto de las directrices también se proyecta en el ámbito normativo. Existe una abultada nueva legislación en la materia de datos que incorpora al mercado ingentes cantidad que antes se encontraban en territorios de mayor protección bajo el manto regulador homogéneo del RGPD. Así:

- Reglamento (UE) 2022/868 del Parlamento Europeo y del Consejo de 30 de mayo de 2022 relativo a la gobernanza europea de datos y por el que se modifica el Reglamento (UE) 2018/1724 (Reglamento de Gobernanza de Datos).

- Reglamento (UE) 2023/2854 del Parlamento Europeo y del Consejo, de 13 de diciembre de 2023, sobre normas armonizadas para un acceso justo a los datos y su utilización, y por el que

se modifican el Reglamento (UE) 2017/2394 y la Directiva (UE) 2020/1828 (Reglamento de Datos).

- Reglamento (UE) 2025/327 del Parlamento Europeo y del Consejo, de 11 de febrero de 2025, relativo al Espacio Europeo de Datos de Salud, y por el que se modifican la Directiva 2011/24/UE y el Reglamento (UE) 2024/2847

Asimismo, el desarrollo de tecnologías emergentes como la inteligencia artificial, "Internet de las cosas", robótica, nanotecnología, biotecnología o neurotecnología presentan nuevos retos que deben ser ponderados en la interpretación de las normas jurídicas a fin de que provean de un ámbito de protección idóneo para la orientación del uso de las tecnologías conforme a la normatividad del dogma de la dignidad humana.

Realizada esta introducción se pasa a razonar porque el Comité Europeo de Protección de Datos asume la claudicación del derecho a la protección de datos frente a las exigencias del mercado.

2. LA VIRTUALIDAD HABILITANTE DEL INTERÉS LEGÍTIMO FRENTE AL CONTENIDO NORMATIVO DEL DERECHO FUNDAMENTAL.

De acuerdo con lo dispuesto en el artículo 8.1 de la Carta de Derechos Fundamentales de la Unión Europea: "*Toda persona tiene derecho a la protección de los datos de carácter personal que la conciernan*". Asimismo, su apartado segundo: "*Estos datos se tratarán de modo leal, para fines concretos y sobre la base del consentimiento de la persona afectada o en virtud de otro fundamento legítimo previsto por la ley. Toda persona tiene derecho a acceder a los datos recogidos que la conciernan y a su rectificación*".

El 6.1f) del RGPD proporciona una base jurídica para el tratamiento de datos personales en la medida en de que "el tratamiento sea necesario para los fines de los intereses legítimos perseguidos por el responsable del tratamiento o por un tercero, salvo que prevalezcan los intereses o los derechos y libertades fundamentales del interesado que requieran la protección de datos personales, en particular cuando el interesado sea un menor".

Las directrices reconocen que la base legal del interés legítimo ex artículo 6.1f) RGPD es la que, como regla general, es utilizada por entidades privadas para el *scraping* para el entrenamiento de IA generativa. Tanto es así que las directrices únicamente van dirigidas hacia las entidades privadas, pues respecto de las públicas, el referido artículo determina que no se aplicará al tratamiento de datos realizado por las autoridades públicas en el ejercicio de sus funciones. En consecuencia, se podría decir la base jurídica del *scraping* sería el interés legítimo por antonomasia.

Y siendo esto así, estas directrices prescinden de las cautelas más acentuadas que plantea el propio Comité Europeo de Protección de Datos en sus Directrices 1/2024, 8 de octubre de 2024, sobre el tratamiento de datos personales basadas en Artículo 6.1 f) del RGPD. En su párrafo noveno se consideraba lo siguiente:

> *Elartículo 6.1f) del RGPD no puede considerarse una base jurídica "por defecto". Por el contrario, antes de recurrir a dicha base jurídica, el responsable del tratamiento debe realizar una evaluación cuidadosa del tratamiento previsto y seguir una metodología específica. El carácter abierto del artículo 6(1)(f) del RGPD8 no significa necesariamente que esta base jurídica deba considerarse como una que solo pueda utilizarse como "último recurso" en situaciones raras e imprevistas, ni que deba considerarse como una última opción si no se aplican otras bases jurídicas. Tampoco debe considerarse el artículo 6(1)(f) como una opción preferida por los responsables del tratamiento, ni su uso debe extenderse indebidamente para eludir requisitos legales específicos o porque se considere menos restrictivo que las demás bases jurídicas del artículo 6(1) del RGPD. En otras palabras, el artículo 6(1)(f) no debe considerarse como una "puerta abierta" para legitimar todas las actividades de tratamiento de datos que no se encuadren en ninguna de las otras bases jurídicas del RGPD. Más bien, conviene recordar que el artículo 6.1 f), al igual que cada uno de los fundamentos jurídicos establecidos en el artículo 6.1 del RGPD, debe interpretarse de forma restrictiva".*

El Comité Europeo de Protección de Datos ante la envergadura que está tomando el desarrollo de la IA generativa no efectúa unas consideraciones generales adecuadas a la complejidad del fenómeno. Efectivamente, recuerda, como lo hiciera entonces, que para que la aplicación del artículo 6.1f) sea lícita, deben cumplirse tres condiciones acumulativas: la persecución de un interés legítimo; la necesidad de tratar los datos para dicho fin; y que los derechos o libertades fundamentales del interesado no prevalezcan sobre dicho interés.

Sin embargo, con la Autoridad Holandesa de Protección de Datos el *scraping* es una técnica que con carácter general atenta frente al contenido esencial del derecho a la protección de datos. De acuerdo la STC español 292/2000, de 30 de noviembre:

> *«7. De todo lo dicho resulta que el contenido del derecho fundamental a la protección de datos consiste en un poder de disposición y de control sobre los datos personales que faculta a la persona para decidir cuáles de esos datos proporcionar a un tercero, sea el Estado o un particular, o cuáles puede este tercero recabar, y que también permite al individuo saber quién posee esos datos personales y para qué, pudiendo oponerse a esa posesión o uso. Estos poderes de disposición y control sobre los datos personales, que constituyen parte del contenido del derecho fundamental a la protección de datos se concretan jurídicamente en la facultad de consentir la recogida, la obtención y el acceso a los datos personales, su posterior almacenamiento y tratamiento, así como su uso o usos posibles, por un tercero, sea el Estado o un particular.*
>
> *Y ese derecho a consentir el conocimiento y el tratamiento, informático o no, de los datos personales, requiere como complementos indispensables, por un lado, la facultad de saber en todo momento quién dispone de esos datos personales y a qué uso los está sometiendo, y, por otro lado, el poder oponerse a esa posesión y usos.*
>
> *En fin, son elementos característicos de la definición constitucional del Derecho fundamental a la protección de datos personales los derechos del afectado a consentir sobre la recogida y uso de sus datos personales y a saber de los mismos.*
>
> *Y resultan indispensables para hacer efectivo ese contenido el reconocimiento del derecho a ser informado de quién posee sus datos personales y con qué fin, y el derecho a poder oponerse a esa posesión y uso requiriendo a quien corresponda que ponga fin a la posesión y empleo de los datos. Es decir, exigiendo del titular del fichero*

que le informe de qué datos posee sobre su persona, accediendo a sus oportunos registros y asientos, y qué destino han tenido, lo que alcanza también a posibles cesionarios; y, en su caso, requerirle para que los rectifique o los cancele.»

Esta jurisprudencia plantea identidad sustancial de fundamento con la mantenida por el Tribunal Constitucional Federal Alemán o el Tribunal Europeo de Derechos Humanos a partir del derecho a la vida privada.

Las directrices, sin perjuicio de algunas cautelas o puntualizaciones, en términos generales, valida la técnica del *scraping* en el entrenamiento de los sistemas de IA. Y esta afirmación es tan obvia que realiza cierto un esfuerzo dialéctico ímprobo en justificar los elementos más cuestionables del *scraping*, pues esta técnica pulveriza los principios dimanantes del derecho a la protección de datos que de acuerdo con el artículo 5.2 del RGPD, que el responsable del tratamiento debe probar que cumple:

a) Lícita, leal y transparente en relación con el interesado («licitud, lealtad y transparencia»)
b) recogidos con fines determinados, explícitos y legítimos, y no serán tratados ulteriormente de manera incompatible con dichos fines
c) Adecuados, pertinentes y limitados a lo necesario en relación con los fines para los que son tratados

En consecuencia, si somete a los principios inherentes al derecho a la protección de datos a una degradación de su carácter normativo para transmutarlo en meras figuras retóricas no cabría considerar que existe una ponderación equilibrada entre el interés de tercero y los derechos y libertades fundamentales de los interesados que habilite la invocación lícita de tal base jurídica que fundamente el tratamiento de datos de carácter personal.

Si el contenido esencial del derecho a la protección de datos radica en "*un poder de disposición y control de datos*" la obligación de la transparencia constituye *conditio sine qua non* del

tratamiento lícito de los datos, pues sin la debida información tiende a ser imposible el ejercicio de esa potestad de disposición y control.

Pues bien, por lo que respecta a la transparencia, las directrices determinan que según cómo se diseñe exactamente el tratamiento, el responsable podría no tener que informar individualmente a los interesados si la provisión de información resulta imposible o requiere un esfuerzo desproporcionado. Efectivamente, el *scraping* se diseña para captar masivamente datos, a priori, de toda índole para el entrenamiento de los modelos de IA. En alguna medida son las redes de enmalle a la deriva -"cortinas de la muerte"- prohibidas por la Unión Europea por tener por resultado una captura indiscriminada de la fauna marina. En estos términos, puede ser realmente complicado poder notificar a los interesados que sus datos están siendo tratados por un tercero, pero esta constatación sin un análisis de mayor profundidad lleva a la validación acrítica de la práctica.

Si acudimos al artículo 14.5 b) del RGPD dicho precepto determina que se podrá excepcionar de la obligación de comunicar el tratamiento cuando:

> *… dicha información resulte imposible o suponga un esfuerzo desproporcionado, en particular para el tratamiento con fines de archivo en interés público, fines de investigación científica o histórica o fines estadísticos, a reserva de las condiciones y garantías indicadas en el artículo 89, apartado 1, o en la medida en que la obligación mencionada en el apartado 1 del presente artículo pueda imposibilitar u obstaculizar gravemente el logro de los objetivos de tal tratamiento. En tales casos, el responsable adoptará medidas adecuadas para proteger los derechos, libertades e intereses legítimos del interesado, inclusive haciendo pública la información..*"

Aun cuando pueda considerarse que los fines de archivo en interés público, de investigación científica o histórica o estadísticos no sean los únicos que habiliten a tal excepción, la norma está estableciendo con tales supuestos un estándar elevado de interés público que habilite tal excepción, lo cual no se puede predicar, sin más, respecto del interés de los titulares de modelos de IA de efectuar su entrenamiento

Además, el RGPD puede incurrir en una patología de "calidad en la ley" en los términos del TEDH. Para este Tribunal cualquier injerencia en los derechos fundamentales -como la intimidad o la libertad- debe estar habilitada por una «ley» suficiente, lo que significa que la norma no solo debe existir formalmente, sino reunir condiciones mínimas de claridad y previsibilidad que garanticen los derechos.

En este caso, no se establece un catálogo cerrado de actividades -que, por tanto, habrá de ser interpretado de manera restrictiva-, pero tampoco se determinar las medidas de garantías que, por otra parte, confía en el agente que acomete tal injerencia (finaliza el precepto "en tales caso, el *responsable adoptará medidas adecuadas para proteger los derechos, libertades e intereses legítimos del interesado…)*

El principio de finalidad también presenta una fuerte conexión con el derecho a la protección de datos como potestad de disposición. De acuerdo con lo dispuesto en el artículo 5.1 b) RGPD, lo datos deben ser "*recogidos con fines determinados, explícitos y legítimos, y no serán tratados ulteriormente de manera incompatible con dichos fines; de acuerdo con el artículo 89, apartado 1, 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 («limitación de la finalidad»)"*

No obstante, de acuerdo con el art. 6.4 RGPD se encuentran permitidos tratamientos secundarios no consentidos siempre que exista una relación concreta lógica y estrecha entre las finalidades y se respeten las legítimas expectativas del interesado -STJUE, de 20 de octubre de 2022, caso Digi (C-77/21)-. Requerimientos que, a priori, no encuentran compatibilidad con las características de la técnica examinada.

Similares consideraciones caben realizarse con el principio de minimización de datos. De acuerdo con este principio los datos personales deben ser adecuados pertinentes y limitados a lo necesario para los fines para los que se procesan. Lecturas laxas o demasiado dúctiles de este principio lo deja reducido a la nada cuando se trata de entrenar a un modelo de propósito general (Chatgpt, Gemini, Copilot, etc), pues los fines son en buena medida indeterminados.

En definitiva, el *scraping* como técnica avanzada de *big data* procesa masivamente datos cuyos titulares no pueden saber qué entidades los están tratando ni con qué finalidad, en consecuencia, enerva el poder de las personas de disponer y controlar sus datos. Como apuntan Solove y Woodrow: "*Esta práctica viola prácticamente todos los principios clave de las leyes de privacidad, incluyendo la equidad, los derechos y el control individuales, la transparencia, el consentimiento, la especificación de la finalidad y las restricciones al uso secundario, la minimización de datos, la transferencia posterior y la seguridad de los datos. La extracción de datos ignora las leyes de protección de datos diseñadas en torno a estos requisitos".* Por esta razón, la Autoridad Holandesa de Protección de Datos (*Autoriteit Persoonsgegevens*, AP) ha dejado claro en sus directrices (*Guidelines for scraping by private individuals and private organisations*) que el web scraping es casi siempre ilegal bajo el RGPD.

En estos términos, no se puede afirmar cabalmente que el derecho a la protección de datos mantenga una vigencia efectiva en el plano de lo fáctico, pues el poder de disposición de los datos, en cuanto contenido esencial del derecho, ha quedado enervado por la realidad digital y, en el plano de lo jurídico, por un exceso de ponderación del interés del mercando en detrimento del ámbito de protección del derecho.

3. LA TRIVIALIZACIÓN DE LAS CATEGORÍAS ESPECIALES DE DATOS

En cuanto a las categorías especiales de datos la carencia del desarrollo de criterios que conduzcan a su protección es absoluta. Se cita la STJUE del caso GC y otros, C-136/17, la cual es poco aplicable para esta técnica, pues se refiere a la obligación del gestor de un motor de búsqueda de estimar una solicitud de retirada de enlaces a una página web en la que figuran datos personales comprendidos en las categorías especiales.

Por supuesto, no existe ninguna referencia a las categorías especiales de datos por inferencias. Se refiere a la deducción de información relativa a esta categoría de datos a partir de otros ordinarios de los cuales se infieren a partir del cálculo algorítmico o la combinación con otros.

Así, por ejemplo, la orientación política de una persona puede ser deducida a partir de sus interacciones en redes sociales, etc.

Esta cuestión tomada en serio debería llevar a una limitación del volumen de datos objeto de tratamiento. La dicción del artículo 9.1 RGPD es suficientemente ilustrativa: "Quedan prohibidos el tratamiento de datos personales que revelen el origen étnico o racial, las opiniones políticas, las convicciones religiosas o filosóficas, o la afiliación sindical, y el tratamiento de datos genéticos, datos biométricos dirigidos a identificar de manera unívoca a una persona física, datos relativos a la salud o datos relativos a la vida sexual o las orientación sexuales de una persona física" -todo ello, sin perjuicio de las excepciones que se relacionan en el apartado segundo-.

No solo se prohíbe el tratamiento de datos que se refieran directamente a tales categorías de datos sino también el "tratamiento de datos personales que revelen" por inferencia tales características de las personas relacionadas con aspectos significativamente íntimos de su personalidad.

4. LA INCONSISTENCIA DE LAS MEDIDAS DE MITIGACIÓN: EL ESPEJISMO DE LA ANONIMIZACIÓN.

La evidencia científica constata que la combinación de grandes conjuntos de datos y el desarrollo de la inteligencia artificial hacen cada vez más difícil garantizar que una persona no pueda ser reidentificada a partir de datos aparentemente anónimos.

En 2000, una investigadora estimó que el 87,1 % de la población estadounidense podía ser identificable mediante la combinación de código postal, sexo y fecha de nacimiento. Para ello se vincularon registros sanitarios supuestamente desidentificados con datos del censo electoral.[3] En 2008, unos investigadores demostraron que a partir de las valoraciones de 500.000 suscriptores de Netflix combinados con información de IMDb podían identificar a los suscriptores de la primera. En 2013 otro grupo analizó durante 15 meses de datos de movilidad de 1,5 millones de personas. Bastaron cuatro puntos espacio-temporales para singularizar al 95

[3] Sweeney, Latanya, *Simple Demographics Often Identify People Uniquely,* Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

% de los individuos. En consecuencia, los patrones de localización poseen una elevada capacidad identificativa[4].

Cuando existen solo varios operadores globales de modelos de lenguaje de gran tamaño en un contexto oligopolísticos global la cuestión adquiere una relevancia sustancial. En el contexto de la IA las posibilidades de reidentificación son exponencialmente mayor habida cuenta de su capacidad de inferencia.

5. HACIA UNA EXPECTATIVA RAZONABLE DE PRIVACIDAD EN EL ENTORNO DIGITAL

Procede efectuar una reflexión en profundidad sobre el equilibrio del desarrollo tecnológico y la privacidad de las personas. Razón por la cual emergió el derecho a la protección de datos.

El camino decidido parece que va en la dirección de deprimir cualquier expectativa razonable de privacidad en el entorno digital, lo cual supone un riesgo efectivo para los derechos individuales, pero también para la democracia.

Frente a técnicas de tanto impacto debe realizarse un escrutinio estricto en el que prevalezca la protección de los derechos y libertades de las personas. A la luz del contenido esencial del derecho a la protección de datos y del RGPD el *scraping web* debe circunscribirse a los fines de interés público, fines de investigación científica e histórica o fines estadísticos

En consecuencia, debe a comenzarse a perfilar un derecho a la privacidad en el entorno digital que favorezca la mayor autodeterminación informativa de las personas y dirija las grandes oportunidades que comparta la IA a un mayor desarrollo humano que es la esencia sobre la que descansa el programa constitucional.

[4] De Montjoye et ale, *Unique in the Crowd: The privacy bounds of human mobility,* Nature, 1376 (2013)