

LOS MERCADOS DE INSUMOS (*INPUTS*) EN EL ENTRENAMIENTO DE LA INTELIGENCIA ARTIFICIAL: EL NUEVO LEVIATÁN DEL DERECHO DE LA COMPETENCIA

INPUT MARKETS IN ARTIFICIAL INTELLIGENCE TRAINING: THE NEW LEVIATHAN OF COMPETITION LAW

Vanessa JIMÉNEZ SERRANÍA*

RESUMEN

Este artículo plantea la dificultad inherente a la definición de un mercado de inteligencia artificial, analizando el «stack» tecnológico como configurador estructural del poder de mercado a través de capas funcionales, entre las que se encuentra la capa de datos de entrenamiento. Se plantea la necesidad de una distinción en esta capa entre datos brutos y *datasets* anotados (datos de entrenamiento que pueden llegar a considerarse esenciales) teniendo en cuenta su diferente relevancia e interés económico en la actualidad, evaluando aspectos vinculados al acceso a los mismos (incluyendo su portabilidad e interoperabilidad). Mediante el análisis de casos recientes, se busca demostrar la importancia de esta distinción taxonómica en la capa de datos del stack, donde los términos *data* y *dataset* no deberían considerarse sinónimos ya que plantean retos y mercados diferenciados que han de ser tenidos en cuenta desde la perspectiva del Derecho de la competencia.

Palabras clave: inteligencia artificial, poder de mercado, datos de entrenamiento, stack de IA, datasets, derecho de la competencia, DMA.

ABSTRACT

This article addresses the inherent difficulty in defining an artificial intelligence market by analyzing the technological «stack» as a structural determinant of market power across functional layers, including the training data layer. It argues for the need to distinguish within this layer between raw data and annotated datasets (i.e., training data that may, in certain circumstances, be considered essential), taking into account their differing contemporary relevance and economic value, and assessing issues related to access, including portability and interoperability. Through the analysis of recent cases, the article seeks to demonstrate the importance of

* Profesora Permanente Laboral de Derecho Mercantil de la Universidad de Salamanca. Dirección de correo electrónico: vjserrania@usal.es. Trabajo realizado en el marco del Proyecto de Investigación «Identidad digital, propiedad sobre activos digitales, propiedad intelectual y mercado ante las tecnologías disruptivas y los nuevos mundos digitales (DIG-IPROPMARKT_3.0)», (PID2023-151076OB-I00), 2024-2027.
Fecha de recepción: 01/05/2026// Fecha de aceptación: 01/06/2026

this taxonomic distinction within the data layer of the stack, where the terms «data» and «dataset» should not be treated as synonymous, as they give rise to distinct challenges and markets that must be considered from a competition law perspective.

Keywords: artificial intelligence, market power, training data, AI stack, datasets, competition law.

SUMARIO: I. REFLEXIONES INICIALES: ¿PODEMOS DEFINIR UN MERCADO DE LA INTELIGENCIA ARTIFICIAL?—II. EL «STACK» DE IA Y LA CONFIGURACIÓN DEL PODER DE MERCADO: CAPAS FUNCIONALES, «INPUTS» ESTRATÉGICOS Y DINÁMICAS COMPETITIVAS: 1. Las «capas» tecnológicas del «stack» de IA; 2. El rol determinante de la capa de los datos.— III. LOS DATOS COMO CAPA DEL STACK Y COMO INSUMO «ESENCIAL»: 1. Algunas precisiones terminológicas preliminares; 2. Mercado de datos: un mercado de insumos informacionales generales: 2.1. Taxonomía de datos: visión general; 2.2. Mercado de datasets para el entrenamiento de la IA, 2.3. Diferencias conceptuales y su relevancia doctrinal; 2.4. Implicaciones para el derecho de la competencia europeo; 2.5. Dimensiones adicionales: portabilidad, interoperabilidad y responsabilidad.— IV. DOS CASOS Y UNA REFLEXIÓN: LA IMPORTANCIA DE LA IDENTIFICACIÓN DEL INPUT CONCRETO EN EL STACK DE DATA: 1. El caso de Google AI vs. editores web y creadores de contenido en Youtube: abuso de posición dominante y control de los datos en el stack de data de IA; 2 El caso SP – Alphabet: la aplicación del art. 6.11: la DMA y el acceso a los datos en el stack de data de IA; 3. Reflexiones en torno al caso LAION: una lectura prospectiva desde el derecho de la competencia: 3.1. Contextualización del caso LAION; 3.2. Reflexión final: la sustancial diferencia entre datos y dataset para el entrenamiento de IA y sus implicaciones en términos de competencia.— V. BIBLIOGRAFÍA.

CONTENTS: I. INITIAL REFLECTIONS: CAN WE DEFINE AN ARTIFICIAL INTELLIGENCE MARKET?— II. THE AI «STACK» AND THE CONFIGURATION OF MARKET POWER: FUNCTIONAL LAYERS, STRATEGIC INPUTS, AND COMPETITIVE DYNAMICS: 1. The Technological «Layers» of the AI Stack; 2. The Determinative Role of the Data Layer.— III. DATA AS A LAYER OF THE STACK AND AS AN «ESSENTIAL» INPUT: 1. Preliminary Terminological Clarifications; 2. The Data Market: A Market for General Informational Inputs: 2.1. Taxonomy of Data: a Global View; 2.2. The Market for AI Training Datasets; 2.3. Conceptual Distinction and Its Doctrinal Relevance; 2.4. Implications for European Competition Law; 2.5. Additional Dimensions: Portability, Interoperability, and Accountability.— IV. TWO CASES AND A REFLECTION: THE IMPORTANCE OF IDENTIFYING THE SPECIFIC INPUT WITHIN THE DATA STACK: 1. The Google AI Case vs. Web Publishers and YouTube Content Creators: Abuse of Dominant Position and Control of Data within the AI Data Stack; 2. The SP – Alphabet Case: The Application of Article 6(11) DMA and Access to Data within the AI Data Stack; 3. Reflections on the LAION Case: A Forward-Looking Competition Law Perspective: 3.1. Contextualization of the LAION Case; 3.2. Final Reflection: The Substantive Difference Between Data and AI Training Datasets and Its Implications for Competition.—V. BIBLIOGRAPHY.

I. REFLEXIONES INICIALES: ¿PODEMOS DEFINIR UN MERCADO DE LA INTELIGENCIA ARTIFICIAL?

La inteligencia artificial (de aquí en adelante, IA) se ha convertido en el epicentro del cambio de nuestro actual paradigma económico y social. Aprender este fenómeno implica, por un lado, comprender una tecnología en constante evolución y diversificación y, por otro, buscar fórmulas que permitan adecuar la realidad jurídica a esta revolución tecnológica que parece imparable.

Desde luego, abordar cualquiera de los dos aspectos mencionados anteriormente no es sencillo ya que somos presos de una carrera cuya meta, en mi opinión, ni siquiera está claramente definida (más allá de intereses económicos inmediatos) y de una ansiedad contextual que nos impide una reflexión pausada y profunda.

A lo anterior se une el hecho de que la IA (y, de manera especialmente significativa, el de la denominada inteligencia artificial generativa) se halla en una fase

de desarrollo exponencial caracterizada por una estructura competitiva incipiente y compleja, elevada volatilidad tecnológica y una densa opacidad informativa. La configuración actual conlleva la coexistencia de múltiples arquitecturas técnicas aún en evolución, la diversidad de modelos¹ y una asimetría informativa estructural entre los operadores del mercado y los observadores regulatorios².

Teniendo en cuenta este contexto, como no podría ser de otro modo, los intentos de definición de mercados de referencia en la IA permanecen en fase de construcción conceptual³. Así, la delimitación de un mercado autónomo de la inteligencia artificial constituye una de las cuestiones más complejas en el análisis contemporáneo del Derecho de la competencia, en la medida en que tensiona los presupuestos clásicos sobre los que se ha construido la noción de mercado relevante. Un aspecto cardinal de esta dificultad radica en determinar si la integración de sistemas de IA altera cualitativamente la naturaleza funcional del producto o servicio hasta el punto de conferirle una autonomía competitiva suficiente para configurar un mercado diferenciado. En efecto, cuando las capacidades de la IA no pueden replicarse (ni siquiera aproximarse) mediante inteligencia humana o recursos convencionales, emerge una categoría técnico-económica distinta que cuestiona la sustituibilidad estricta entre prestaciones tradicionales y automatizadas. Capacidades como el procesamiento masivo de datos en tiempos ínfimos, la operación ininterrumpida sin degradación o la adopción de decisiones autónomas generan ventajas estructurales que pueden justificar la segmentación de mercados previamente integrados.

¹ Uno de los ejemplos claros respecto a esta situación son los conocidos como modelos fundacionales. Como señalan Schrepel y Pentland, en el momento actual coexisten modelos fundacionales generales, de redes y personales siendo desarrollados y ofrecidos bajo condiciones diferentes, desde modelos «cerrados» a modelos «open source». SCHREPEL y PENTLAND (2025), págs. 1086 y 1087. En el mismo sentido, *vid.* CARUGATI (2023B). Por otro lado, recordemos que dentro de la categoría general de IA quedan comprendidos los sistemas de machine learning, de redes neuronales y de deep learning. Sin ánimo de exhaustividad, encontramos (i) algoritmos de aprendizaje supervisado — que pueden ser clasificados como Regresión logística, árboles de decisión, bosques aleatorios, máquinas de vectores de soporte (SVM), Naive Bayes (Bayes ingenuo), vecinos más cercanos (KNN), redes neuronales (por ejemplo, perceptrón multicapa) y clasificadores de aumento de gradiente (por ejemplo, XGBoost, LightGBM)—; (ii) algoritmos de aprendizaje no supervisado —donde encontramos algoritmos de clustering—, (tales como K-Means, agrupamiento jerárquico, DBSCAN, modelos de mezcla gaussiana —GMM—), algoritmos de reducción de dimensionalidad (como son los de análisis de componentes principales (PCA), incrustación estocástica de vecinos distribuidos t (t-SNE), aproximación y proyección de variedad uniforme —UMAP—) y algoritmos de aprendizaje de reglas de asociación (como Apriori, FP-Growth); (iii) algoritmos de aprendizaje por refuerzo —como son Q-Learning—, redes neuronales profundas Q (DQN), gradientes de políticas (por ejemplo, REINFORCE), optimización proximal de políticas (PPO), gradientes de políticas deterministas profundas (DDPG)-; (iv) algoritmos de aprendizaje profundo (*deep learning*) —entre los que podemos citar, las redes neuronales convolucionales (CNN), las redes neuronales recurrentes (RNN), las redes de memoria a largo y corto plazo (LSTM), las unidades recurrentes con compuertas (GRU), los transformadores, las redes generativas antagónicas (GAN), los autoencoders y las máquinas de Boltzmann profundas (DBM)—; (v) los algoritmos que comprenden métodos de conjunto — como son de ensacado (bagging) con bosques aleatorios, potenciación (boosting) con AdaBoost, aumento de gradiente (XGBoost, LightGBM, CatBoost) y apilamiento (stacking); y, por último encontramos, (vi) otros algoritmos de inteligencia artificial — como son los algoritmos genéticos, el «recocido» (annealing simulado) y algoritmos de inteligencia de enjambre (por ejemplo, optimización por enjambre de partículas). Para una mejor comprensión de cada uno de ellos y una clasificación más completa ver: <https://medium.com/@anil.jain.baba/comprehensive-guide-to-machine-learning-deep-learning-and-ai-algorithms-3d96f5a0c077>.

² Como ha sido señalado por varios autores, existe todavía un gran margen de mejora en la manera en la que las distintas autoridades, especialmente en el ámbito de competencia, comprenden y se aproximan a la IA (especialmente en el caso de la GenAI). *Vid.* ZIERMANN, (2025), pág. 4, especialmente nota 38; RADIC y STOUT (2024), págs. 130-131; SCHREPEL y PENTLAND (2025), págs. 1093 y 1095.

³ Sobre esta cuestión, ver ZIERMANN, Fabian (2025), págs. 354-358; SCHREPEL y PENTLAND (2025), págs. 1085-1086; JIMÉNEZ SERRANÍA (2026), págs. 145-172.

Sin embargo, esta posible autonomía no se manifiesta de forma uniforme ni automática, lo que explica la dificultad de definir un mercado de la IA como categoría independiente. En algunos supuestos, la inteligencia artificial da lugar a bienes y servicios cualitativamente inéditos, cuya funcionalidad carece de paralelismo con productos preexistentes, como ocurre con los sistemas de análisis automatizado de documentación jurídica o científica en contextos de gran volumen, donde la combinación de velocidad, exhaustividad y capacidad adaptativa excede objetivamente las posibilidades humanas o de algoritmos convencionales⁴. En tales casos, la innovación tecnológica no se limita a extender mercados existentes, sino que puede configurar mercados nuevos. No obstante, la consolidación de estos nuevos mercados depende, evidentemente, de la emergencia de una demanda específica asociada a prestaciones que antes eran imposibles o económicamente inviables, lo que introduce un elemento contingente y evolutivo en la delimitación del mercado.

A esta complejidad se añade la necesidad de adoptar un enfoque dinámico en la definición del mercado relevante. Allí donde la IA introduce capacidades que transforman la naturaleza económica de la oferta o de la demanda, el análisis no puede limitarse a una fotografía estática de la sustituibilidad actual, sino que debe incorporar la trayectoria previsible de los procesos de sustitución tecnológica⁵. De este modo, el mercado relevante deja de ser una herramienta meramente clasificatoria para convertirse en un instrumento de comprensión evolutiva de las dinámicas competitivas, especialmente en entornos caracterizados por innovación acelerada y convergencia tecnológica⁶.

Este panorama general se enmaraña más aún si consideramos la relación entre inteligencia artificial e inteligencia humana como frontera última de la sustituibilidad. El análisis ya no se limita a comparar productos o servicios técnicamente análogos, sino que debe evaluar si determinadas funciones realizadas por sistemas de IA compiten directamente con actividades tradicionalmente desempeñadas por personas. Esta ampliación del marco comparativo introduce una complejidad singular, pues la inteligencia humana pasa a operar como referencia competitiva. Históricamente, se ha presumido la superioridad estructural de las capacidades humanas en ámbitos que requieren juicio contextual, creatividad o interacción social compleja, lo que situaba a la IA como complemento y no como sustituto⁷. Sin embargo, el desarrollo reciente de modelos funda-

⁴ THIEMANN (2025), págs.11 y 12.

⁵ KERBER (2023), págs. 12-15.

⁶ Drex1 y Kim, partiendo de las teorías de Hayek y de Kerber, argumentan que la competencia no es solo un mecanismo de asignación de recursos, sino un sistema para generar y difundir conocimiento mediante la experimentación paralela. En el ámbito de la IA, esto implica que el derecho de la competencia no debe intentar «diseñar» el mercado favoreciendo un modelo de negocio específico (como el código abierto frente al propietario), ya que la evidencia empírica sobre cuál es superior para la innovación es inconcluyente. Por consiguiente, la política antitrust debe centrarse en salvaguardar la integridad del proceso de experimentación y la autonomía estratégica de los desarrolladores. Esto es especialmente crítico ante la configuración del *stack* de IA, donde la dependencia de los pequeños innovadores respecto a insumos críticos controlados por firmas de «Big Tech» (como la capacidad de cómputo masiva y grandes conjuntos de datos) puede sofocar la diversidad tecnológica y la libertad de elección necesaria. DREXL y KIM (2026), págs. 181-218.

⁷ ACEMOGLU y RESTREPO (2019), págs. 3-30. Encontramos evidencias de esta sustitución, por ejemplo, en sectores como la fotografía de «stock», donde se han documentado caídas del 20 por 100 en la demanda de imágenes no generadas por IA. *Vid.* PEUKERT, (2025), págs.17 y 18.

cionales basados en técnicas de *deep learning* ha comenzado a erosionar esta frontera, reduciendo el espacio de funciones exclusivamente humanas⁸.

Podría sostenerse que la principal dificultad a la hora de configurar un mercado autónomo de la inteligencia artificial reside en su naturaleza intrínsecamente híbrida, dinámica y transversal, que desborda los parámetros clásicos de delimitación del mercado relevante en el Derecho de la competencia de la Unión Europea. A diferencia de otros sectores en los que los productos o servicios pueden ser identificados a partir de criterios relativamente estables de sustituibilidad desde la perspectiva de la demanda (tal como tradicionalmente ha recogido la Comunicación de la Comisión sobre la definición de mercado relevante⁹), la inteligencia artificial se presenta como un conjunto de tecnologías y capacidades cuya función económica varía sustancialmente en función de su grado de integración y de los usos a los que se destina¹⁰. Así, puede operar como un insumo intermedio integrado en mercados digitales ya existentes (mejorando funcionalidades de búsqueda, recomendación o automatización) o, en otros casos, dar lugar a nuevos espacios competitivos con identidad propia, como sucede con determinados servicios basados en modelos fundacionales.

Entre estos dos extremos se sitúa, sin embargo, una amplia zona intermedia caracterizada por la inestabilidad de sus contornos y por su fuerte dependencia de factores tecnológicos (como el grado de generalidad y reutilización de los modelos), económicos (incluidas las economías de escala, alcance y aprendizaje) y funcionales (en particular, los distintos casos de uso y contextos de despliegue). Esta indeterminación plantea retos evidentes para la aplicación de los instrumentos tradicionales de análisis competitivo, en la medida en que las relaciones de sustituibilidad no siempre pueden captarse adecuadamente atendiendo exclusivamente al producto final, ni mediante una aplicación estricta del test del monopolista hipotético en su formulación convencional.

La delimitación del mercado relevante no puede agotarse en el nivel de las aplicaciones o servicios finales, sino que exige una aproximación más estructural que tenga en cuenta la organización interna de la cadena de valor que permite la creación y explotación de sistemas de inteligencia artificial. En consecuencia, la definición de mercado no puede apoyarse en categorías tradicionales como los «productos finales», sino que debe incorporar una comprensión

⁸ Los modelos de lenguaje de gran escala constituyen un ejemplo paradigmático de esta transformación, al haber irrumpido en sectores intensivos en trabajo cognitivo —como el periodismo, la traducción o la documentación jurídica— ofreciendo prestaciones equiparables o incluso superiores en términos de velocidad, consistencia y capacidad de procesamiento, aunque con limitaciones en aspectos contextuales o éticos. Desde la perspectiva del análisis antitrust, esta sustitución funcional, incluso si es parcial, puede ser suficiente para redefinir la extensión del mercado e integrar la oferta algorítmica como competidor efectivo de la oferta humana. No obstante, esta convergencia no permite formular generalizaciones: la sustituibilidad debe evaluarse de forma casuística, atendiendo a la función económica concreta, al contexto competitivo y a la elasticidad de la demanda frente a soluciones humanas o automatizadas.

⁹ Comunicación de la Comisión relativa a la definición de mercado de referencia a efectos de la normativa de la Unión en materia de competencia, C/2023/6789, párr. 25 a 31, https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=OJ:C_202401645.

¹⁰ No obstante, los potenciales problemas competitivos en el ámbito de la inteligencia artificial no son, ni mucho menos, ajenos a la Comisión Europea. De hecho, el mismo 2024, la Comisión Europea publicó un «Policy Brief» titulado «*Competition in Generative AI and Virtual Worlds*» que examinaba de manera exhaustiva los retos competitivos inherentes a la inteligencia artificial generativa (IA generativa), con especial atención a las barreras de entrada y las dinámicas de mercado a lo largo de su cadena de valor. *Vid.* KOWALSKI, VOLPIN y ZOMBORI (2024), págs.2-10.

funcional del sistema tecnológico, atendiendo a las relaciones de complementariedad, dependencia y convertibilidad entre modelos, infraestructuras y capas de servicio¹¹.

Abundando en este razonamiento y en línea con la evolución reciente de la práctica decisoria de la Comisión en mercados digitales (particularmente en ámbitos como las plataformas en línea, la publicidad digital o los ecosistemas tecnológicos), resulta necesario desagregar el ecosistema de la IA en sus distintas capas funcionales: infraestructura computacional, acceso y control de datos, desarrollo de modelos fundacionales, herramientas de desarrollo y despliegue, y aplicaciones finales. Ahora bien, estas capas de la cadena de valor no deben entenderse únicamente como niveles técnicos, sino como ámbitos de provisión y control de «inputs» para el desarrollo y la comercialización de sistemas de IA. En particular, la capacidad computacional, los conjuntos de datos de alta calidad, los modelos fundacionales y determinadas herramientas de desarrollo constituyen insumos estratégicos cuya disponibilidad, coste y condiciones de acceso pueden condicionar decisivamente la competencia en los niveles superiores del ecosistema.

Desde esta perspectiva, el análisis de lo que se denomina en terminología tecnológica, «stack» de IA¹² se revela no solo metodológicamente adecuado, sino imprescindible desde el punto de vista del Derecho de la competencia de la UE. Esta desagregación permite identificar con mayor precisión posibles mercados relevantes en cada nivel, evaluar la existencia de poder de mercado —incluyendo situaciones de control o estrangulamiento de «inputs» clave— y comprender los mecanismos mediante los cuales dicho poder puede construirse, reforzarse o trasladarse entre capas. En particular, fenómenos como la integración vertical, el acceso privilegiado a insumos críticos (como datos o capacidad computacional), o la generación de efectos de red y de aprendizaje pueden dar lugar a posiciones de ventaja que no resultan visibles si el análisis se limita al producto final.

Por consiguiente, el examen detallado del «stack» de IA constituye el punto de partida necesario para evaluar las dinámicas competitivas específicas de la inteligencia artificial.

II. EL «STACK» DE IA Y LA CONFIGURACIÓN DEL PODER DE MERCADO: CAPAS FUNCIONALES, «INPUTS» ESTRATÉGICOS Y DINÁMICAS COMPETITIVAS

Por stack debe entenderse la arquitectura estratificada e interdependiente (la «pila tecnológica» o *technology stack* dentro de la cadena de suministro de IA o *AI supply chain*) que soporta el despliegue operativo de sistemas de inteligencia artificial¹³. Esta estructura integra capas tecnológicas funcionalmente solida-

¹¹ Como señalan Radic y Stout, la IA no es una tecnología única y monolítica, *vid.* RADIC y STOUT (2024), págs. 108 y 109.

¹² *Vid.* ALAILY (2024), pág. 6.

¹³ OCDE (2025), *Competition in artificial intelligence infrastructure*, OCDE Roundtables on Competition Policy Papers, N° 330, p. 9, https://www.oecd.org/en/publications/competition-in-artificial-intelligence-infrastructure_623d1874-en/full-report/component-5.html. Además del Policy Brief de la Comisión Europea y los distintos estudios de la OCDE que señalamos en este texto, estos aspectos son también señalados por

rias: desde la infraestructura física de hardware especializado (aceleradores de IA como GPUs y TPUs, semiconductores y centros de datos) hasta los servicios de software en la nube (XaaS), pasando por flujos de datos, modelos fundacionales, aplicaciones *downstream* y canales de distribución.

Como veremos, esta estratificación trasciende la mera yuxtaposición técnica, distinta de la cadena de suministro física (centrada en la producción lineal de componentes), para configurarse como un sistema de dependencias recíprocas, en el que la funcionalidad, rendimiento y control de cada estrato condicionan estructuralmente los restantes. Esta indisoluble interrelación entre capas constituye su rasgo definitorio en clave jurídica¹⁴.

Desde la perspectiva del Derecho de la competencia, la determinación del *stack* de IA permite visualizar los potenciales mapas estructurales del poder de mercado, permitiendo identificar nodos críticos de control y dinámicas de integración vertical que proyectan efectos exclusionarios aguas arriba y aguas abajo. De hecho, la concentración sustancial tanto en el poder computacional como en los modelos fundacionales genera auténticas estructuras de cuello de botella que restringen la contestabilidad del ecosistema¹⁵.

Por consiguiente, esta cadena de valor de la inteligencia artificial presenta una heterogeneidad estructural que se proyecta desde los niveles más básicos de provisión de hardware —en particular, semiconductores avanzados y unidades de procesamiento de alto rendimiento (GPU/TPU)— hasta los estratos aplicativos en los que se concretan los servicios finales. En los primeros eslabones, la demanda se caracteriza por su marcada transversalidad: los mismos insumos computacionales y arquitecturas técnicas son utilizados en múltiples sectores (desde el entretenimiento digital hasta la investigación científica o las infraestructuras *blockchain*), lo que impide establecer una correspondencia unívoca entre mercados de insumos y mercados de producto.

En los niveles superiores del «stack» (o pila) de IA, la organización de la actividad económica viene determinada en gran medida por la decisión estratégica de «*make or buy*», esto es, por la elección entre desarrollar internamente

los estudios recientes de distintas autoridades de la competencia, destacando los estudios de las autoridades inglesa (COMPETITION AND MARKETS AUTHORITY (2024), AI Foundation Models, <https://www.gov.uk/government/publications/ai-foundation-models-update-paper>), francesa (AUTORITÉ DE LA CONCURRENCE (2024), Generative artificial intelligence: the Autorité issues its opinion on the competitive functioning of the sector, <https://www.autoritedelaconcurrence.fr/en/press-release/generative-artificial-intelligence-autorite-issues-its-opinion-competitive>) y canadiense (COMPETITION BUREAU CANADA (2024), Artificial intelligence and competition. Discussion Paper, Competition Bureau Canada, <https://competition-bureau.canada.ca/sites/default/files/documents/AICompetition-Discussion-Paper-240320-ver3-e.pdf>).

¹⁴ Un ejemplo de estas interacciones lo encontramos en la arquitectura CUDA de NVIDIA. CUDA (Compute Unified Device Architecture) es la plataforma de cómputo paralelo desarrollada por NVIDIA, que permite programar las GPUs para tareas generales más allá del renderizado gráfico tradicional. CUDA transforma las GPUs en aceleradores masivos para IA, simulaciones científicas y análisis de datos, ejecutando miles de hilos simultáneamente mediante extensiones de C/C++ y bibliotecas optimizadas. En el contexto jurídico que mencionamos su interrelación con capas superiores del *stack* de IA (*frameworks* como PyTorch, bibliotecas cuDNN) genera dependencias críticas que plantean cuestiones regulatorias sobre competencia y soberanía tecnológica. Vid. <https://www.nvidia.com/es-es/solutions/ai-factories/>. Es decir, esta tecnología no funciona aislada: conecta de forma inseparable el hardware (las propias GPUs), el software base (CUDA), las bibliotecas especializadas y los *frameworks* de desarrollo (Vid. <https://medium.com/@tahirbalarabe2/what-is-nvidia-cuda-4013e5f4deb5>). Esa interdependencia entre capas es precisamente lo que define su importancia jurídica, ya que crea una dependencia técnica respecto a NVIDIA que plantea serios retos.

¹⁵ STUCKE y EZRACHI, (2025), págs. 161-168.

capacidades críticas (incluyendo acceso a datos, entrenamiento de modelos o infraestructura) o recurrir a proveedores externos mediante servicios en la nube, APIs o modelos preentrenados. Esta elección no es neutral desde la perspectiva del Derecho de la competencia, en la medida en que condiciona el grado de dependencia respecto de determinados operadores y puede consolidar posiciones de poder en los niveles intermedios de la cadena de valor.

En este contexto, la interacción entre decisiones de integración vertical, control de insumos (inputs) estratégicos y condiciones de acceso a los mismos incide directamente en la configuración de las relaciones de sustituibilidad y, en última instancia, en la delimitación del mercado relevante. La coexistencia de estructuras integradas con modelos basados en el suministro de terceros genera una arquitectura compleja en la que las fronteras entre mercados aguas arriba y aguas abajo se vuelven difusas y contingentes. Como resultado, la identificación de mercados relevantes claramente diferenciados se ve dificultada, exigiendo un análisis que tenga en cuenta no solo las características de los productos finales, sino también las dependencias técnicas, económicas y organizativas que articulan el conjunto del «stack» de IA.

Con todo, antes de abordar el análisis detallado de estas dinámicas, resulta imprescindible delimitar conceptualmente qué se entiende por «stack» de IA e identificar sus principales capas funcionales para, posteriormente, poder abordar ciertos aspectos relativos a los insumos («inputs») que consideramos relevantes.

1. Las «capas» tecnológicas del «stack» de IA

Como indicábamos anteriormente, el «stack» de IA comprende, como su propio nombre indica, distintas capas de diferentes tecnologías y elementos que se encuentran, en la mayoría de los casos, interrelacionados de una manera indisoluble.

Encontramos, por tanto, una primera capa de hardware, en la que encontramos tanto los chips —semiconductores— como los centros de almacenamiento de datos; una segunda, que podemos denominar XaaS, integrando todos los servicios de software (incluyendo cloud) que son necesarios para la implementación de estas tecnologías; una tercera fundamental y diversa, que sería la de los datos, sobre la que profundizaremos más adelante; una cuarta cuya complejidad hemos mencionado anteriormente, que es la de los modelos fundacionales utilizados para el entrenamiento y, posteriormente, las capas vinculadas con la propia explotación específica de la IA, integradas por las aplicaciones, los distintos canales de distribución¹⁶.

¹⁶ El informe de la OCDE sobre «Competition in Artificial Intelligence Infrastructure» (2025) resalta la estructura del «stack» de IA como un ecosistema interdependiente de capas físicas y digitales altamente concentradas, donde destacan los chips especializados (aceleradores de computación) con cadenas de suministro globales limitadas por la dependencia de tecnologías críticas como la litografía ultravioleta extrema (EUV) producida solo por ASML, los centros de datos como «fábricas de IA» que concentran entrenamiento e inferencia con barreras por costos de energía, enfriamiento y conectividad, y los servicios de «cloud» dominados por hiperescaladores como AWS, Azure y Google Cloud, que fomentan integración vertical y economías de escala. No aparece en el mismo una capa específica dentro de la «infraestructura de suministro» relativa a los datos (aunque sí aparece una relativa a los data center). En nuestra opinión, consideramos que esta capa es una de las capas esenciales por lo que la incluimos en nuestra clasificación y la estudiamos más detalladamente en el presente artículo OCDE (2025), págs. 9-16.

Cada una de estas capas reviste una problemática particular y merece un estudio detallado a la hora de abordar una aproximación al tema que nos ocupa. Sin ánimo de exhaustividad, volveremos a citar la importante concentración en la oferta de la primera¹⁷ y segunda capas mencionadas¹⁸, debido tanto a la escasez de materias primas como de recursos para poder hacer frente a los desembolsos que este tipo de elementos supone, así como a las complejas cadenas de valor que llevan aparejadas¹⁹. No menos compleja resulta la capa de los datos. La base de cualquier sistema de inteligencia artificial radica en los datos de entrenamiento, que se clasifican principalmente en estructurados —altamente organizados y consultables en bases de datos, como hojas de cálculo con filas y columnas— y no estructurados, sin formato predefinido, como textos, imágenes, vídeos, audio o código. Estos datos pueden provenir de fuentes diversas: públicas (scraping web, conjuntos como Common Crawl, ImageNet o LibriSpeech), privadas (como Google Search o YouTube), generados por usuarios en redes sociales o generados sintéticamente mediante técnicas como GANs o *autoencoders* variacionales para aumentar muestras, preservar privacidad y mitigar sesgos o desequilibrios²⁰. La calidad, cantidad, diversidad y representatividad de estos datos determinan directamente el rendimiento del modelo, siguiendo el principio «*garbage in, garbage out*», ya que un conjunto bien curado minimiza sesgos, mejora la generalización y permite captar patrones complejos²¹. Pese a la pluralidad de fuentes, en la práctica nos encontramos con un panorama en el que se observa una tendencia a la monopolización de estos datos²².

Dejando al margen los distintos modelos fundacionales ya mencionados anteriormente, si ascendemos a las capas de «aplicación» nos encontramos que el despliegue y utilización de aplicaciones basadas en estos requieren, por regla general, el uso de herramientas especializadas, así como de sistemas de seguridad que garanticen un empleo responsable de la tecnología. Entre dichas herramientas se incluyen las destinadas a la orquestación de procesos (como Apache Airflow, Prefect o Databricks Workflows), que permiten automatizar y coordi-

¹⁷ Vid. HUGGINS, JOHNSTON, MUNDAY y XU (2023), pág. 534.

¹⁸ Vid. CARUGATI (2023), págs. 2-9.

¹⁹ Recordemos que, por ejemplo, los ecosistemas de *cloud* van a estar caracterizados por una compleja cadena de valor que está compuesta por varios niveles interconectados, pudiendo identificarse los siguientes actores: a) Los proveedores de infraestructura física que son las empresas responsables de la construcción, gestión y mantenimiento de los centros de datos y la infraestructura física subyacente (servidores, redes, almacenamiento, energía, refrigeración, etc.). Estos actores requieren inversiones de capital muy elevadas y una gestión eficiente de recursos, lo que limita la cantidad de competidores a escala global; b) Los proveedores de servicios de nube (hiperescaladores) como son, por ejemplo, Amazon Web Services, Microsoft Azure, Google Cloud. Estos actores ofrecen servicios IaaS, PaaS y SaaS a escala global, y suelen estar integrados verticalmente, abarcando desde la infraestructura hasta servicios y aplicaciones avanzadas. Su dimensión e implantación les permite beneficiarse de economías de escala y de red; c) Los desarrolladores de aplicaciones y servicios que utilizan las plataformas y recursos ofrecidos por los hiperescaladores para crear aplicaciones y servicios especializados (por ejemplo, software de gestión empresarial, análisis de datos, inteligencia artificial). Estos desarrolladores pueden ser empresas independientes o filiales de los propios hiperescaladores; d) Los integradores, consultores y revendedores que son intermediarios cuya labor es ayudar a los clientes finales a seleccionar, implementar y gestionar soluciones en la nube, facilitando la migración, integración y personalización de servicios. Su papel es clave para reducir la complejidad técnica y contractual, especialmente para pymes y organizaciones con menos recursos tecnológicos; y, por último, e) los clientes finales que comprenden los desarrolladores de IA. OECD, *Competition in the Provision of Cloud Computing Services*, OECD Publishing, Paris, 2025, págs. 14-16.

²⁰ A mayor abundamiento sobre esta cuestión, ver GÁLVEZ DURÁN (2025), págs. 8-18.

²¹ RADIC y STOUT, (2024), pág. 115.

²² DANG (2025), págs. 120 a 127.

nar flujos de trabajo, gestionar infraestructuras de datos y optimizar la eficiencia del proceso de aprendizaje automático²³.

Por otro lado, las aplicaciones prácticas de estos modelos son diversas e incluyen asistentes conversacionales, herramientas empresariales y programas de codificación. En algunos casos, un único modelo fundacional resulta suficiente para ofrecer la experiencia de usuario deseada; en otros, se combinan varios modelos a fin de obtener resultados más precisos y robustos. Por último, la distribución de estas aplicaciones va a poder revestir diversos usos y entornos (desde usos meramente internos a usos más comerciales), lo que hará que encontremos despliegues tanto en *cloud* como perimetrales (*edge*) o locales (*on-premises*). Como vemos la heterogeneidad y las potenciales interdependencias funcionales entre las distintas capas del «stack» de IA son importantes y habrán de ser tenidas en cuenta en el estudio concurrencial.

2. El rol determinante de la capa de los datos

Como ya hemos visto anteriormente, los modelos de IA presentan grados sustantivos de diferenciación técnica y funcional, determinantes para su análisis desde la perspectiva del mercado relevante. De manera simplificada podríamos distinguir entre modelos generalistas (esto es, modelos entrenados sobre conjuntos de datos masivos, amplios y polivalentes —como grandes corpus textuales o multimodales—) y modelos de dominio específico, entrenados sobre *datasets* restringidos, etiquetados y orientados a finalidades profesionales o sectoriales concretas (por ejemplo, en los ámbitos financiero, médico o jurídico). Esta distinción no es meramente técnica, sino que responde a la naturaleza misma de los inputs de entrenamiento: la composición, volumen, calidad y diversidad de los datos determinan el rendimiento, la generalización y el propósito del modelo resultante, planteando interrogantes cruciales sobre su sustituibilidad empírica que se abordarán en los siguientes apartados.

Explicaremos esto con un ejemplo ilustrativo: un modelo como Stable Diffusion, entrenado sobre miles de millones de imágenes anotadas, no puede considerarse sustituto funcional de otro como GPT-4, orientado al procesamiento de lenguaje natural, o de Llama en tareas de código, dado que la heterogeneidad de los *datasets* de entrenamiento —imágenes vs. texto vs. código— genera capacidades intrínsecamente no intercambiables. Como destaca la OCDE en su análisis de la infraestructura de IA²⁴, los datos de entrenamiento constituyen un input crítico y no replicable en la cadena de valor, cuya escasez o concentración puede generar barreras de entrada y posiciones dominantes en capas específicas del «stack» de AI, exacerbando la diferenciación entre modelos.

Estas diferencias se amplían si se atiende al propio ecosistema de uso de los sistemas de IA, que no se agota en el componente algorítmico, sino que abarca infraestructuras computacionales, interfaces, dispositivos de interacción y entornos de implementación. En particular, la OCDE enfatiza que la dependencia de *datasets* de alta calidad —a menudo sectoriales o propietarios— limita la

²³ ALAILY (2024), pág. 8.

²⁴ OCDE (2025), «Mapping relevant data collection mechanisms for AI training», págs. 9-16.

interoperabilidad y la sustituibilidad, convirtiendo el control sobre estos inputs en un factor clave para la configuración de mercados relevantes diferenciados.

Esta heterogeneidad funcional de los modelos de IA, determinada en gran medida por la composición cualitativa y cuantitativa de los *datasets* de entrenamiento, evidencia la centralidad de los mercados de datos como capa estratégica del «stack» de IA. En efecto, los datos no solo constituyen el input primario para el desarrollo de capacidades diferenciadas, sino que configuran un ámbito competitivo autónomo cuya estructura —caracterizada por efectos de red, barreras de acceso y concentración en posesión— incide directamente en la delimitación del mercado relevante aguas abajo. Frente a esta realidad, resulta imprescindible analizar los mercados de datos en su especificidad, examinando sus dinámicas competitivas propias y su proyección sobre las condiciones de competencia en el ecosistema de la inteligencia artificial.

III. LOS DATOS COMO CAPA DEL STACK Y COMO INSUMO «ESENCIAL»

1. Algunas precisiones terminológicas preliminares

La capa de datos en el «stack» de IA emerge como un insumo estratégico determinante de la competencia en inteligencia artificial, configurando no solo el rendimiento técnico de los modelos sino también las dinámicas de poder de mercado aguas abajo. No obstante, antes de entrar a estudiar y proponer ciertos aspectos para la reflexión sobre esta capa nos parece necesario hacer una breve distinción conceptual precisa entre inputs en sentido amplio, datos («data») y *datasets* estructurados.

Desde un punto de vista técnico, podríamos decir que el término input abarca toda entrada informacional, paramétrica u operativa que un sistema de IA recibe para cumplir una función específica, integrando las fases de entrenamiento, inferencia y mejora iterativa del modelo. Esta definición abarcaría tanto los datos personales y no personales como las instrucciones humanas (incluidos prompts), los parámetros algorítmicos, las señales sintéticas o los flujos procedentes de sensores o APIs.

Doctrinalmente, este input trasciende la condición de mera materia prima técnica para constituirse en presupuesto funcional de la decisión algorítmica, determinando la trayectoria inferencial del sistema y la calidad de sus outputs. La OCDE refuerza esta caracterización al identificar los inputs de entrenamiento —datos diversos y de alta calidad— como insumos críticos cuya escasez relativa o composición sesgada genera no solo riesgos técnicos, sino también barreras competitivas estructurales en el «stack» de AI.

En este «stack» de AI, los datos constituyen una capa estratégica diferenciada que, a nuestro entender, exige distinguir conceptualmente entre *data* y *dataset*²⁵. Por *data* debe entenderse el conjunto difuso de información (individual o agregada) que sirve de sustrato material a los sistemas de IA, independientemente de su estructuración, procedencia o calidad, abarcando tanto datos

²⁵ Vid. <https://www.databricks.com/blog/what-is-dataset>; <https://medium.com/@Bright-Data/what-is-a-dataset-definition-use-cases-benefits-and-example-9aaf5ecc301e>

estructurados (bases relacionales) como no estructurados (texto, imágenes, audio, código fuente).

Su relevancia jurídica radica en su condición de insumo esencial para el aprendizaje automático (regido por el axioma *garbage in, garbage out*), pues datos sesgados, incompletos o no representativos generan modelos con rendimientos deficientes y riesgos jurídicos asociados. El *dataset*, en contraste, configura una agregación estructurada y teleológicamente orientada de data, resultante de decisiones humanas sobre selección, balanceo, anotación y preprocesamiento para fines específicos de entrenamiento, validación o auditoría.

Esta distinción resulta pivotal desde el Derecho de la competencia: el *dataset* incorpora una dimensión teleológica y de gobernanza que lo convierte en unidad autónoma de análisis, susceptible de imputación responsable, auditoría (sesgo, exactitud) y trazabilidad, mientras la *data* suele circular en flujos opacos y automatizados²⁶.

2. Mercado de datos: un mercado de insumos informacionales generales

En un sentido amplio, el mercado de datos se configura como un entorno en el que se articula la oferta, licencia, reutilización y monetización de datos considerados como bien económico, con independencia del conjunto concreto sobre el que se extienden. En esta perspectiva, el dato se trata como un input genérico para la producción de modelos algorítmicos, servicios de recomendación, desarrollos de IA generativa o sistemas de toma de decisiones automatizada. La idea central es que el valor no se asocia únicamente a la información contenida, sino a su capacidad de entrenar, validar o afinar modelos de aprendizaje automático, con lo que el mercado de datos se entiende como un mercado de insumos digitales que precede a la producción de bienes y servicios finales.

2.1. Taxonomía de datos para entrenamiento de IA desde la perspectiva europea

Desde la óptica económica, el mercado de datos en la Unión se caracteriza por la existencia de múltiples categorías de datos (por ejemplo, datos de navegación, de consumo, sensoriales, de geolocalización, de imágenes etiquetadas, de textos estructurados), que pueden circular bajo distintas modalidades de licencia, acceso restringido o condiciones de uso. En este contexto, la competencia se entiende de forma relativamente amplia: un agente puede competir en un determinado servicio de IA aunque no disponga del mismo *dataset* que un rival, siempre que pueda acceder a otros datos de similar calidad y volumen, o bien generarlos indirectamente mediante técnicas de aumento de datos, simulación o recopilación alternativa.

No obstante, para poder dimensionar esta cuestión nos parece relevante intentar establecer una especie de taxonomía de los datos que pueden ser utilizados para el entrenamiento de sistemas de IA ya que, como ya ha sido apuntado

²⁶ OCDE (2025), «Mapping relevant data collection mechanisms for AI training», pp.8-11 y 18.

por parte de la OCDE, de su naturaleza jurídica derivan importantes consecuencias respecto a la competencia efectiva en el mercado²⁷:

Desde la perspectiva jurídica europea, la distinción primordial entre datos personales y no personales se fundamenta en el art. 4.1 del RGPD²⁸, que define datos personales como «toda información sobre una persona física identificada o identificable (“el interesado”); se considerará persona física identificable toda persona cuya identidad pueda determinarse, directa o indirectamente, en particular mediante un identificador, como por ejemplo un nombre, un número de identificación, datos de localización, un identificador en línea o uno o varios elementos propios de la identidad física, fisiológica, genética, psíquica, económica, cultural o social de dicha persona»²⁹. En contraste, los «datos no personales» incluyen información no vinculada a individuos identificables (tales como datos industriales, meteorológicos, o agregados y anonimizados)³⁰.

Podemos establecer otra distinción entre datos «públicos» y datos «privados», vinculándolos a la gobernanza de datos en la Unión Europea, donde se determinan regímenes jurídicos diferenciados de acceso y reutilización para el entrenamiento de sistemas de IA. Los datos públicos, esto es, los generados en el sector público, están destinados a la reutilización abierta, conforme a la Directiva (UE) 2019/1024 sobre datos abiertos y reutilización de la información del sector público. Este acceso se refuerza con el Reglamento de Gobernanza de Datos³¹, el Reglamento de Datos³² y, más recientemente y de manera sectorial, con el Reglamento sobre el Espacio Europeo de Datos de Salud³³. Por el contrario, los datos privados, originados en empresas de sectores como la banca o la industria farmacéutica, se rigen por regímenes contractuales y de propiedad intelectual, caracterizados por licencias propietarias y cláusulas de exclusividad. Aunque el Reglamento de Datos establece en su art. V una obligación excepcional de cesión a entidades públicas, el régimen ordinario preserva el control comercial de los titulares privados³⁴.

²⁷ OCDE (2025), «Mapping relevant data collection mechanisms for AI training», pág.19.

²⁸ Reglamento (UE) 2016/679, de 27 de abril, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos).

²⁹ Como señala Alba Ribera, el desarrollo de la inteligencia artificial (IA) generativa presenta un desafío jurídico fundamental debido a su dependencia de vastos conjuntos de datos no estructurados, frecuentemente obtenidos mediante técnicas de *web scraping*, que integran de forma inevitable datos de carácter personal en los modelos. Esta arquitectura técnica genera riesgos de privacidad significativos, como la posible filtración de información sensible a través de la «memorización» de datos o ataques maliciosos de inversión. Desde la perspectiva del RGPD, la licitud del tratamiento para el entrenamiento de estos modelos es compleja; mientras que el consentimiento resulta inmanejable a escala masiva, el «interés legítimo» se ha consolidado como la base legal más viable para los desarrolladores, siempre que se superen los juicios de necesidad y proporcionalidad exigidos por las autoridades de control. RIBERA MARTÍNEZ (2025), págs.10-12.

³⁰ Además de estos datos, Picht señala también una categoría diferenciada de las anteriores que son los datos seudonimizados (art. 4.5 del RGPD). Estos se diferencian de los datos anonimizados en que todavía permiten la identificación del interesado si se dispone de información adicional (por ejemplo, mediante claves de asignación). Por ello, están sujetos a las disposiciones del RGPD. PICT (2025), pág.15.

³¹ Reglamento (UE) 2022/868 del Parlamento Europeo y del Consejo, de 30 de mayo, relativo a la gobernanza europea de datos y por el que se modifica el Reglamento (UE) 2018/1724.

³² Reglamento (UE) 2023/2854, de 13 de diciembre, 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.

³³ Reglamento (UE) 2025/327 del Parlamento Europeo y del Consejo, de 11 de febrero, 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.

³⁴ Hemos de puntualizar que el Reglamento (UE) 2025/327 impone a las entidades privadas (prestadores sanitarios, titulares de datos de salud electrónicos) obligaciones clave para uso primario y secundario entre las

Por otro lado, encontramos una clasificación derivada de la práctica publicitaria que nos parece interesante y que distingue entre datos de primera parte, segunda parte y tercera parte, lo que permite afinar la comprensión de la cadena de valor de datos desde una perspectiva de mercado. Los datos de primera parte serían aquellos que el propio proveedor de IA recoge directamente de sus usuarios, generando un flujo de retroalimentación que mejora progresivamente la calidad del modelo y refuerza su posición competitiva. La acumulación de datos de primera parte, combinada con la ³⁵limitación de interoperabilidad y de portabilidad, puede consolidar posiciones de poder sostenidas, especialmente en mercados de servicios de IA generalistas dominados por unos pocos agentes globales. Los datos de segunda parte proceden de socios comerciales directos (por ejemplo, bancos, *retailers*, fabricantes) y se intercambian en un entorno contractual relativamente cerrado; en este caso podemos encontrar acuerdos de suministro de datos que pueden contener cláusulas de exclusividad o de no compartición que limiten la capacidad de terceros de acceder a flujos de datos similares. Finalmente, los datos de tercera parte se adquieren a intermediarios especializados, tales como proveedores de mercados de conjuntos de datos o plataformas de datos (donde también se integrarían los proveedores de *datasets* de entrenamiento que analizaremos en el apartado siguiente).

Desde la perspectiva de la estructura técnica del dato, la distinción más relevante opera entre datos textuales, datos visuales y de audio, y datos numéricos, sensoriales y contextuales. Los datos textuales se subdividen en texto libre y texto estructurado. Por otro lado, hemos de señalar una diferencia importante entre el que podríamos denominar dato «en bruto» y el *dataset* estructurado. El primer tipo de datos, procedente de *web scraping*, libros, artículos, foros o redes sociales, se emplea en el entrenamiento de modelos de lenguaje generalistas o multilingües. El *dataset* estructurado, en cambio, incluye estos datos anteriores marcados y anotados, aumentando, por consiguiente, su valor en el mercado. En estos casos podemos llegar incluso a plantear la dependencia de terceros respecto a estos *datasets* estructurados y etiquetados, especialmente cuando la replicación de esa estructura y de esa anotación supone un costo desproporcionado o una imposibilidad práctica.

que figuran las de proporcionar a los datos que recaben. Recordemos que este Reglamento considera como tenedor de datos de salud a «toda persona física o jurídica, autoridad pública, agencia u otro organismo del sector de la asistencia sanitaria o el asistencial, incluidos los servicios de reembolso cuando sea necesario, así como cualquier persona física o jurídica que desarrolle productos o servicios destinados al sector de la salud, el de la asistencia sanitaria o el asistencial, que desarrolle o cree aplicaciones de bienestar, que efectúe investigaciones científicas relacionadas con el sector de la asistencia sanitaria o el asistencial o que actúe como registro de mortalidad, así como las instituciones, órganos u organismos de la Unión, que tengan bien: i) el derecho o la obligación, de conformidad con el Derecho de la Unión o nacional aplicable y en su calidad de responsable o corresponsable del tratamiento, de tratar datos de salud electrónicos personales para fines de prestación de asistencia sanitaria o fines asistenciales, o para fines de salud pública, reembolso, investigación, innovación, formulación de políticas, estadísticas oficiales o seguridad del paciente o para fines de regulación, o ii) la capacidad de poner a disposición datos de salud electrónicos no personales mediante el control del diseño técnico de un producto y de los servicios conexos, incluido mediante el registro, la entrega, la limitación del acceso o el intercambio de dichos datos» (Cfr. art. 2.1. t). No obstante, en el propio texto se establece que no pretende obstaculizar ni sustituir los acuerdos contractuales u otros existentes (Considerando 52); así como que «*las medidas legales, organizativas y técnicas para preservar los derechos de propiedad intelectual e industrial o los secretos comerciales podrían incluir acuerdos contractuales comunes de acceso a datos de salud electrónicos, obligaciones específicas en el permiso de datos en relación con dichos derechos (...)*» (Considerando 60).

³⁵ Vid. <https://amperity.com/blog/first-party-vs-third-party-data-what-marketers-need-to-know-in-2026>

La función económica del dato permite articular, finalmente, mercados de datos de entrenamiento, de validación, de prueba y de uso. Los datos de entrenamiento pueden estar comprendidos por grandes conjuntos sin etiquetar o etiquetados (luego volveremos sobre este aspecto al tratar los *datasets*). La negativa de acceso a estos conjuntos, o la imposición de condiciones desfavorables, puede afectar a la capacidad de las nuevas entrantes de entrenar modelos de calidad comparable, tanto en mercados de IA generalista como de IA de dominio específico. Los datos de validación y prueba se utilizan para evaluar el rendimiento de los modelos, configurando un mercado de datos de evaluación, donde la competencia vigila la posible configuración de ciertos conjuntos de evaluación como estándares de facto, que puedan favorecer a determinados modelos frente a alternativas de mercado. En el ámbito de los datos de uso, la competencia observa cómo la retroalimentación de usuarios refuerza la posición del modelo dominante mediante la acumulación de datos de uso de alta calidad, generando ventajas de escala y experiencia que se transmiten a otros servicios de IA asociados, en línea con la preocupación de la Comisión sobre la capacidad de los grandes actores de reforzar su posición a través de la reutilización de datos.

Complementariamente, podemos mencionar datos de dominio específico, como, por ejemplo, los datos médicos que pueden comprender imágenes, historiales clínicos, datos de ensayos clínicos u otros conjuntos de datos sanitarios, cuyo acceso puede ser indispensable para competir en sistemas de diagnóstico asistido o de soporte clínico algorítmico. También podríamos poner como ejemplo los datos en el sector legal o el financiero.

Por último, es obligado mencionar a los conocidos como «datos sintéticos»³⁶. Desde el punto de vista jurídico, los datos sintéticos se definen como conjuntos de datos generados por modelos algorítmicos que reproducen la estructura estadística, distribución y características de uno o varios conjuntos de datos reales, de manera que simulan registros de personas, empresas, transacciones o comportamientos sin contener, en principio, datos personales directos. La calidad de un dataset sintético depende en última instancia de la calidad, amplitud y representatividad de los datos reales sobre los que se ha entrenado el modelo generador, de forma que la accesibilidad a esos datos originales se convierte en un factor clave para la capacidad de generar sintéticos de alto rendimiento. Desde la perspectiva económica, la generación de datos sintéticos requiere una inversión significativa en infraestructura de cómputo, algoritmos de simulación y procesos de enriquecimiento estadístico, de forma que no todas las empresas pueden generar datos sintéticos de la misma calidad o escala que los grandes actores tecnológicos.

2.2. Mercado de datasets para el entrenamiento de IA

Como planteábamos anteriormente, en nuestra opinión, dentro de estas diferentes categorías de datos, encontramos una tipología que puede considerarse con entidad autónoma para constituir su propio mercado: los *datasets* para el entrenamiento de la IA. Este mercado habría de definirse de forma mucho más concreta, centrando la atención en el dataset como conjunto estructurado, deli-

³⁶ DANG (2025), págs. 120 a 127.

mitado y con propiedades específicas³⁷. Un *dataset* no es, pues, un dato aislado, sino una colección organizada de registros —estructurados o no—, con características de origen, volumen, periodicidad, formato, calidad, etiquetado y grado de anonimización que determinan su aptitud para entrenar, validar o evaluar un modelo concreto³⁸. En numerosas aplicaciones de IA, el rendimiento del modelo depende directamente de la calidad y especificidad de los *datasets* sobre los que se ha entrenado: por ejemplo, un *dataset* de imágenes médicas etiquetadas de forma muy precisa será decisivo para la fiabilidad de un sistema de diagnóstico asistido por IA.

Como señala Picht, en su informe «Data licensing and data-related provisions in technology transfer agreements», los datos de entrenamiento en inteligencia artificial consisten en conjuntos de información (*datasets*) previamente seleccionados y organizados con el fin de permitir que los modelos de aprendizaje automático identifiquen patrones, adopten decisiones o formulen predicciones. En particular, los sistemas basados en aprendizaje automático —especialmente en los enfoques de aprendizaje supervisado y por refuerzo— dependen de grandes volúmenes de datos para las fases de entrenamiento, validación y ajuste del modelo. Estos conjuntos pueden integrar datos estructurados o no estructurados, procedentes tanto de fuentes humanas como de sistemas automatizados, y tener origen público o privado. Además, se presentan en múltiples formatos, como imágenes o tablas de datos. Así, en el ámbito del reconocimiento de imágenes, es habitual que los datos de entrenamiento consistan en archivos visuales acompañados de etiquetas que identifican los objetos representados³⁹.

Un rasgo distintivo de los datos de entrenamiento es su elevado grado de elaboración (aspecto que los diferencia de los datos en bruto). Este proceso implica tanto la selección de información relevante como su estructuración deliberada para hacerla apta para el aprendizaje automático. Dicho tratamiento resulta fundamental para que el modelo pueda aprender de forma eficaz y extrapolar adecuadamente a situaciones o datos no previamente observados.

Desde la óptica antitrust, el paso de «datos» a «*dataset*» para el entrenamiento de IA supone un cambio de nivel de análisis: la sustitución ya no se razona sobre la posibilidad de acceder a datos genéricos, sino sobre la existencia o ausencia de *datasets* comparables para una tarea específica. Esto introduce una dimensión de indispensabilidad técnica: un *dataset* puede resultar indispensable para entrenar un modelo con cierto nivel de precisión, o para competir en un nicho específico de mercado (por ejemplo, vehículos autónomos, IA de radiología, sistemas de recomendación especializados). En este contexto, un mercado de *datasets* puede delimitarse como un sector donde se negocian conjuntos muy

³⁷ Según el Data Catalog Vocabulary (DCAT) en su versión de 2020 se define como «A collection of data, published or curated by a single agent, and available for access or download in one or more formats.» <https://www.w3.org/TR/vocab-dcat-1/#class-dataset>. Como señala Picht, citando a varios autores, los datos sintéticos son aquellos que se generan mediante modelos matemáticos o algoritmos diseñados específicamente para abordar uno o varios problemas propios de la ciencia de datos. La síntesis de datos consiste en transformar datos originales —frecuentemente de carácter personal— en una representación artificial que, en principio, no permite la identificación de individuos. Este tipo de datos resulta especialmente útil para la creación de conjuntos de datos que preservan la privacidad, al reproducir patrones del mundo real sin exponer información sensible, lo que explica su amplia utilización en sectores como el financiero o el sanitario para el entrenamiento de modelos de inteligencia artificial. PICT (2026), pág. 16.

³⁸ LIU, LUO, TANG, *et al.* (2024), pág. 760.

³⁹ PICT (2025), pág.14.

específicos, cuya escasez o concentración puede generar posiciones de poder significativas.

Hemos de tener en cuenta que, en muchos casos, la sustitución perfecta entre *datasets* no existe, ya que la replicación de un conjunto concreto (por ejemplo, datos de localización de alta frecuencia, conversaciones históricas en un idioma minoritario, registros clínicos acumulados durante décadas) puede ser extremadamente costosa o incluso imposible de obtener por terceros. Esto abre la puerta a configurar ciertos *datasets* como bienes de acceso esencial, cuya negativa de acceso sin justificación objetiva podría constituir abuso de posición dominante bajo el art. 102 TFUE, siempre que se acrediten las condiciones doctrinales clásicas (posibilidad de eliminación de competencia, falta de compensación, ausencia de justificación técnica o económica)⁴⁰.

Así, por ejemplo, los *datasets* de calidad, necesarios para el entrenamiento de la IA médica, de seguridad, geoespacial, financiera o industrial son costosos de replicar. La negativa de acceso a *datasets* de imágenes médicas, sateliticas o industriales puede configurar un problema de acceso esencial. Los datos de audio y habla, estructurados o no estructurados, se utilizan en sistemas de reconocimiento de voz, síntesis de voz o análisis de conversaciones, donde la concentración de grandes *datasets* de voz —especialmente en idiomas minoritarios o con gran diversidad de acentos puede generar ventajas sostenidas para el titular, limitando la capacidad de competidores de desarrollar asistentes de voz o sistemas de análisis de voz comparables.

2.3. Diferencia conceptual y su relevancia doctrinal

De lo anterior podríamos colegir que mientras el mercado de datos amplio que puede ser utilizado para la IA se concibe como un sistema de mercado de insumos informacionales generales, el mercado de *datasets* para el entrenamiento de IA se entiende como un segmento de mercado de conjuntos específicos y estructurados, cuya sustitución debe valorarse en términos de equivalencia funcional para una tarea algorítmica concreta. Esta distinción permite refinar el análisis de poder de mercado y de exclusión: en un mercado amplio de datos, puede existir competencia en múltiples fuentes; en un mercado de *datasets* para el entrenamiento de IA muy especializados, la concentración sobre un reducido número de conjuntos puede generar una dependencia estructural de terceros.

⁴⁰ La aplicación de la doctrina de las facilidades esenciales en la Unión Europea ha experimentado un cambio radical tras la sentencia del TJUE en el caso *Android Auto* (C-233/23, 25 febrero 2025, ECLI:EU:C:2025:110), que abandona el estándar estricto de indispensabilidad del caso «Bronner» (1998) para plataformas digitales abiertas. En el litigio, Enel X Italia denunció a Google por denegar la interoperabilidad de su app «JuicePass» (para la recarga de vehículos eléctricos) con «Android Auto», plataforma de Google para aplicaciones móviles en vehículos; la Autorità Garante della Concorrenza e del Mercato multó a Alphabet con 102M€ por abuso de posición dominante (art.102 TFUE). Recordemos que tradicionalmente la aplicación de la doctrina de las facilidades esenciales exigía indispensabilidad absoluta en el mercado descendente. El TJUE restringe este test a infraestructuras desarrolladas «exclusivamente para uso propio» del dominante, reservadas sin apertura prevista. Bajo la nueva interpretación, si la plataforma (como era el caso de «Android Auto») se diseña «abierto a terceros por naturaleza», la negativa a interoperabilidad constituye abuso incluso sin indispensabilidad. Las únicas salvaguardas restantes son justificaciones objetivas (imposibilidad técnica, seguridad), diluyendo libertad contractual de dominantes en digital. *Vid.* COLANGELO (2025), págs. 1449-1476. En el contexto de los *datasets* para entrenamiento de IA, esta evolución jurisprudencial podría obligar a compartir conjuntos únicos como «facilidades convenientes» si son abiertos/reutilizables, configurándolos como infraestructuras esenciales en mercados nicho sin alternativas razonables.

Como hemos visto, el mercado de datos tiende a vincularse con la gobernanza general del acceso, la interoperabilidad y las políticas de datos abiertos (por ejemplo, en el marco de la Data Governance Act y del Data Act), donde el objetivo es expandir la disponibilidad de datos y facilitar la reutilización de datos públicos y privados. Aquí, la competencia opera de forma indirecta, mediante la creación de infraestructuras (como espacios de datos sectoriales) que reducen barreras de entrada y favorecen la pluralidad de proveedores de datos. En cambio, el mercado de *datasets* para el entrenamiento de IA, en nuestra opinión, se conecta más directamente con el derecho de la competencia sustantivo, en particular con el concepto de acceso esencial y con la evaluación de conductas excluyentes (denegación de acceso, condiciones desfavorables, cláusulas restrictivas en licencias) que puedan afectar a la capacidad de un tercero para entrenar o mejorar sus propios modelos.

Esta dualidad se vuelve aún más relevante en el contexto de la IA generativa y de los grandes modelos de lenguaje. En estos entornos, la calidad del *dataset* de entrenamiento no es un factor marginal, sino determinante para la competitividad del modelo: la posesión de *datasets* de gran volumen y alta calidad puede crear ventajas de costes y de desempeño difíciles de replicar. En este escenario, la delimitación de un mercado de *datasets* de entrenamiento para un determinado tipo de IA (por ejemplo, IA para atención sanitaria, IA para crédito financiero, IA para servicios de transporte) puede resultar crucial para evaluar si una empresa dominante utiliza su posición en datos para reforzar su posición en el mercado de servicios de IA, en claro paralelo con la jurisprudencia clásica sobre infraestructuras esenciales y servicios de red.

2.4. Implicaciones para el derecho de la competencia europeo

En el ordenamiento de la UE, la distinción entre mercados de datos y mercados de *datasets* para el entrenamiento tiene consecuencias directas sobre la aplicación del art. 102 TFUE y sobre la interacción con la nueva regulación digital así como en el control de concentraciones. En este contexto, la noción de *dataset* para el entrenamiento se vuelve clave: es el *dataset* concreto, con sus características específicas, el que debe valorarse como posible bien de acceso esencial, más que «los datos» en abstracto.

En la práctica, esto implica que la prueba de exclusión debe ser más fina: no basta con mostrar que un tercero tiene acceso a otros datos, sino que pueda competir en igualdad de condiciones con el mismo nivel de calidad y especificidad de datos. Si un *dataset* determinado (por ejemplo, un conjunto de datos de tráfico en tiempo real acumulados durante años por un operador dominante) no puede ser replicado razonablemente por terceros, y negarse a compartirlo permite al titular reforzar su posición en el mercado de servicios de navegación, la Comisión podría considerar obligaciones de acceso, siempre que se acrediten las condiciones habituales de exclusión y ausencia de justificación⁴¹.

⁴¹ Como señalan Akman, Fumagalli y Mota en su artículo sobre el Borrador de Directrices de la Comisión Europea sobre abusos de exclusión, estas destacan el papel central de los «accesos esenciales» (insumos, infraestructuras o plataformas clave) como puerta de entrada al mercado para competidores y usuarios. Subrayan que el riesgo de exclusión anticompetitiva aumenta cuando el dominante cuenta con ventajas de incumbencia, primero en el mercado o control de un activo crucial (aunque no necesariamente «indispensable» en el

Además, la distinción resulta relevante para la articulación entre la DMA y el derecho de la competencia clásico. Los gatekeepers, bajo la DMA, están sujetos a obligaciones específicas de acceso a datos (por ejemplo, a datos de actividad y de uso de los servicios), que pueden comprender tanto conjuntos de datos específicos como flujos de datos generales. Podríamos considerar que la negativa de acceso a determinados *datasets* por parte de un gatekeeper puede ser examinada simultáneamente bajo el art. 102 y la DMA⁴² (e incluso sobre la DSA⁴³), según la naturaleza de los datos y del servicio implicado. En este sentido, el mercado de *datasets* para el entrenamiento de IA aparece como el ámbito de intersección más delicado entre regulación *ex ante* y análisis de competencia *ex post*⁴⁴.

Por último, en los últimos meses hemos visto un desarrollo exponencial de empresas que se dedican a la «producción» de estos *datasets* para el entrenamiento (empresas de etiquetado de datos para el entrenamiento de sistemas de IA). Las adquisiciones o tomas de control de grandes plataformas sobre este tipo de empresas plantean riesgos serios para la competencia porque pueden reforzar el acceso exclusivo a insumos críticos, elevar los costes de cambio de los rivales y generar incentivos a discriminar o degradar el servicio a clientes competidores⁴⁵ pero, en muchos casos, quedan fuera del control de las autoridades de competencia.

sentido estricto de «Bronner»). Así, la teoría económica ha identificado múltiples escenarios en que una empresa verticalmente integrada puede usar su control sobre un insumo clave para bloquear o dificultar el acceso de rivales (tanto mediante negativas directas de suministro como mediante limitaciones de acceso (restricciones de acceso, *margin squeeze*, auto-preferencia, etc.), que afectan la capacidad de competidores de usar esa infraestructura o input esencial para competir. Por otro lado, coincidimos en la crítica al recurso por parte de la Comisión a presunciones y criterios fragmentados (las Directrices tratan de forma diferente distintas prácticas que pueden tener efectos de exclusión similares) que pueden dificultar en la práctica el acceso de competidores igualmente eficientes a insumos esenciales. Frente a esta situación los autores proponen que el marco normativo incorpore de forma más clara la protección de esos accesos como núcleo de la prohibición de abuso, vinculando mejor la doctrina del «competidor igualmente eficiente» con la necesidad de garantizar el acceso para competir por precio, calidad e innovación. AKMAN, FUMAGALLI y MOTTA (2025), págs. 234-243.

⁴² Reglamento (UE) 2022/1925 del Parlamento Europeo y del Consejo, de 14 de septiembre, sobre mercados digitales contestables y justos (Reglamento de Mercados Digitales o DMA).

⁴³ Reglamento (UE) 2022/2065 del Parlamento Europeo y del Consejo, de 19 de octubre, sobre un mercado único de servicios digitales y por el que se modifica la Directiva 2000/31/CE (Reglamento de Servicios Digitales o DSA).

⁴⁴ Tanto la DMA como la DSA introducen numerosas obligaciones de acceso a datos y transparencia para supervisores, investigadores y terceros. En el magnífico informe sobre el acceso a datos y algoritmos del CERRE se identifican cincuenta y cuatro (54) obligaciones de acceso a datos y transparencia, clasificándolas en cuatro categorías: 1) acceso de reguladores a datos privados; 2) acceso de partes privadas a datos de otras partes privadas; 3) acceso del público general a ciertos datos; y 4) transparencia de los propios reguladores. En relación con los *datasets*, el estudio se centra en tres casos concretos: 1) el acceso a las bases de datos de publicidad online (Art. 39 DSA), que deben poner a disposición del público, vía API y portal web, información sobre todos los anuncios de pago mostrados; 2) el acceso a datos para investigadores previamente verificados (Art. 40 DSA), según el cual las plataformas de gran tamaño deben compartir *datasets* relevantes para el estudio de riesgos sistémicos, incluyendo datos sobre algoritmos, datos de entrenamiento e incluso códigos, bajo condiciones de seguridad y anonimización; y 3) el intercambio de datos de clics y búsquedas con motores de búsqueda competidores (Art. 6.11 DMA), que obliga a compartir *datasets* de consultas y clics con el fin de aumentar la contestabilidad del mercado. Como se señala en este informe, la implementación de estos accesos requiere un equilibrio cuidadoso entre objetivos de regulación, protección de datos personales, secreto comercial y seguridad, por lo que se deben privilegiar principios de minimización de datos y de acceso lo menos intrusivo posible, así como el uso de medidas técnicas (anónimos, datos sintéticos, salas de datos seguras, APIs restringidas) para hacer viable el acceso a los *datasets* sin comprometer excesivamente los intereses protegidos. EDELSON, GRAEF y LANCIERI (2023).

⁴⁵ Un ejemplo muy ilustrativo es la inversión de Meta en Scale AI —valorada en torno a 14.3/14.8 mil millones de dólares y con una participación del 49 por 100—, que provocó que Google rompiera su re-

2.5. Dimensiones adicionales: portabilidad, interoperabilidad y responsabilidad

Más allá de la dimensión exclusivamente competitiva, la distinción entre mercados de datos y mercados de *datasets* para el entrenamiento de IA tiene implicaciones sobre la portabilidad de datos y la interoperabilidad. Desde la óptica del mercado de datos, la portabilidad favorece la movilidad de insumos; desde la óptica del mercado de *datasets*, puede facilitar la creación de conjuntos alternativos o la mejora de la calidad de los mismos.

Asimismo, la interoperabilidad se configura como un factor clave para la sustitución de *datasets*: si los datos pueden integrarse fácilmente en distintos modelos y plataformas, la dependencia de un *dataset* específico se reduce. En este sentido, la regulación de interoperabilidad, presente en la DMA, puede atenuar el poder de mercado derivado de la posesión de determinados *datasets*. Sin embargo, cuando la interoperabilidad no es suficiente —por ejemplo, porque los datos están altamente estructurados o porque la calidad del etiquetado es irreproducible—, la posibilidad de exclusión persiste y el análisis de competencia debe ser más exigente.

En suma, la diferencia entre mercado de datos y mercado de *datasets* para la IA no es meramente técnica, sino que afecta al núcleo mismo del análisis jurídico-económico de los mercados digitales. Mientras el primero se aproxima a un mercado de insumos generales, el segundo se entiende como un mercado de conjuntos específicos, cuya sustitución debe valorarse en términos de equivalencia funcional para una tarea de modelización concreta. En el marco de la competencia europea, esto permite un enfoque más preciso sobre la indispensabilidad, la exclusión y el acceso esencial, conectando de forma directa la regulación de datos con la doctrina antitrust clásica y con las nuevas herramientas de la DMA y la Estrategia de Datos de la UE⁴⁶.

lación comercial con Scale y que otros clientes como OpenAI o xAI reconsideraran sus vínculos por el temor a que un competidor accediera a información sensible sobre sus datos, prioridades de entrenamiento y prototipos además de sembrar la preocupación respecto al futuro del sector de IA. Vid. <https://www.somo.nl/the-next-big-power-grab-in-ai-why-metas-scale-ai-deal-must-be-stopped/>

⁴⁶ Encontramos un ejemplo reciente en los procedimientos relativos a las obligaciones de interoperabilidad y acceso a datos que inició la Comisión Europea el 27 de enero de 2026 contra Google. Como decíamos, el primer conjunto de procedimientos (DMA.100220) se centra en la obligación de Google, establecida en el art. 6.7 de la DMA, de proporcionar a los desarrolladores de terceros una interoperabilidad gratuita y efectiva con las funcionalidades de hardware y software controladas por el sistema operativo Android de Google. Los procedimientos iniciados se centran en funcionalidades utilizadas por los propios servicios de IA de Google, como Gemini. La Comisión pretende especificar cómo Google debe conceder a los proveedores de servicios de IA de terceros un acceso igualmente efectivo a las mismas funcionalidades disponibles para sus propios servicios. El segundo conjunto de procedimientos (DMA.100209), que analizaremos más en detalle en el epígrafe IV, se refiere a la obligación de Google, prevista en el art. 6.11 de la DMA, de conceder a los proveedores terceros de motores de búsqueda en línea acceso a datos anonimizados de clasificación, consultas, clics y visualizaciones que posee Google Search, en condiciones justas, razonables y no discriminatorias («FRAND»). Estos procedimientos se centran en el alcance de los datos, el método de anonimización, las condiciones de acceso y la elegibilidad de los proveedores de chatbots de IA para acceder a dichos datos. Un cumplimiento efectivo y el acceso a un conjunto de datos útil permitirán a los proveedores terceros de motores de búsqueda optimizar sus servicios y ofrecer a los usuarios alternativas reales a Google Search. (Vid. https://digital-markets-act.ec.europa.eu/commission-opens-proceedings-assist-google-complying-interoperability-and-online-search-data-sharing-2026-01-27_en). Justo antes de finalizar la revisión de este art., el 27 de abril de 2026, la Comisión envió sus conclusiones provisionales respecto al primero de los casos (DMA.100220) describiendo el proyecto de medidas que Google debería aplicar para garantizar que terceros tengan acceso efectivo e interoperabilidad con capacidades clave de Android. Las medidas propuestas tienen por objeto asegurar que los servicios de

IV. DOS CASOS Y UNA REFLEXIÓN: LA IMPORTANCIA DE LA IDENTIFICACIÓN DEL INPUT CONCRETO EN EL STACK DE DATA

1. El caso de Google AI vs. editores web y creadores de contenido en Youtube: abuso de posición dominante y control de los datos en el stack de data de IA

En el marco de la evolución reciente de la vigilancia del mercado de la inteligencia artificial, el 9 de diciembre de 2025 la Comisión Europea adoptó una decisión de inicio de procedimiento antitrust contra Google y Alphabet en el expediente AT.40983 (“Google AI & Data-related practices”)⁴⁷, en el seno del régimen sancionador de la Unión Europea regulado por el art. 11, apartado 6, del Reglamento del Consejo 1/2003 y el art. 2, apartado 1, del Reglamento de la Comisión 773/2004, por indicios de infracción del art. 102 del Tratado de Funcionamiento de la Unión Europea (TFUE) y del art. 54 del Acuerdo sobre el Espacio Económico Europeo. La decisión de apertura de procedimiento, que se extiende a todo el territorio del EEE, no constituye una calificación definitiva de infracción, sino el señalamiento de que el expediente será tratado como prioritario dentro del ordenamiento de competencia de la UE.

La conducta investigada juega en el cruce entre el mercado de los datos y el mercado de los servicios de IA generativa, y se centra en dos prácticas concate-nadas. En primer lugar, Google emplea contenido de terceros editores publicado en la web para generar *AI Overviews* y resúmenes en modo IA en el servicio Google Search, sin ofrecer a los citados editores la posibilidad de negar tal uso del contenido sin perder el acceso a Google Search ni percibir una compensación adecuada por su utilización. En segundo término, la Comisión examina la práctica de Google de recopilar y utilizar videoclips de creadores de contenidos subidos a YouTube para entrenar modelos de IA generativa, igualmente sin facilitar a los creadores el derecho efectivo de rechazo sin pérdida de acceso a la plataforma YouTube y sin la contraprestación equivalente al valor económico del dato generado.

Desde la perspectiva del mercado de los datos, ambas prácticas ponen de relieve la jerarquía funcional que el *stack de IA* configura entre capas de infraestructura, de datos y de servicios: Google Search y YouTube no se limitan a operar como mercados de distribución o de acceso a contenidos, sino que se configuran como plataformas de captura y reutilización de conjuntos de datos en tiempo real, en los que el mismo conjunto de datos puede ser empleado simul-

IA competidores puedan interactuar de forma efectiva con las aplicaciones de los dispositivos Android de los usuarios y ejecutar tareas en consecuencia, como enviar un correo electrónico utilizando la aplicación de correo preferida por el usuario, pedir comida o compartir una foto con amigos. Asimismo, las medidas propuestas permitirán que los proveedores competidores de servicios de IA impulsen la innovación y ofrezcan experiencias de IA profundamente integradas a los usuarios en teléfonos y tabletas Android, junto con los propios servicios de IA de Alphabet, como Gemini. La apertura del acceso a estas capacidades proporcionará a los usuarios de Android en toda la Unión Europea una oferta más amplia de servicios de IA. Para garantizar que estas medidas propuestas resulten efectivas, la Comisión ha invitado a las partes interesadas a presentar observaciones sobre ellas mediante una consulta pública abierta hasta el 13 de mayo de 2026. (Vid. https://ec.europa.eu/commission/presscorner/detail/en/ip_26_887).

⁴⁷ Vid. <https://competition-cases.ec.europa.eu/cases/AT.40983>

táneamente para la mejora de servicios de búsqueda, la generación de productos de IA y, potencialmente, el reentrenamiento de modelos de lenguaje o visión.

La estructura de la conducta investigada revela la centralidad de los datos de uso de terceros como insumo esencial en el modelo de negocio de IA generativa de Google. Lejos de limitarse a la explotación de datos propios o agregados, la Comisión señala la utilización de contenido editorial y de videoclips de creadores, que se integran en flujos de entrenamiento de modelos sin garantizar opciones reales de exclusión o remuneración. Esta configuración abre la posibilidad de que el control de accesos de primera parte a datos de alta calidad —especialmente contenido de profesionales del periodismo, del entretenimiento y de la educación— se traduzca en ventajas de escala, experiencia y rendimiento que restan competitividad a terceros que no dispongan de flujos de datos de similar alcance y representatividad.

Bajo la lógica del art. 102 TFUE, este planteamiento permite articular la hipótesis de una conducta de reutilización asimétrica de datos de terceros, que podría reforzar la posición dominante de Google en el mercado de búsqueda o en el de plataformas de vídeos, al tiempo que se despliega un *feedback loop* en el que los datos de uso se reinyectan de forma estructurada en los modelos de IA, amplificando el coste de entrada de competidores que deben construir *datasets* de entrenamiento a partir de un universo de contenido más acotado o de menor calidad sobre los datos. Vemos que en este caso, el foco se pone en el acceso a los datos «brutos» o con un procesamiento mínimo, no a los *datasets* de entrenamiento.

2. El caso SP – Alphabet: la aplicación del art. 6.11: la DMA y el acceso a los datos en el stack de data de IA

Como mencionábamos anteriormente, la Comisión Europea el 27 de enero de 2026 inició sendos procedimientos contra Google centrados en el incumplimiento de ciertas obligaciones de interoperabilidad (DMA.100220) y de acceso a datos establecidas por la DMA (DMA.100209).

El 16 de abril de 2026 la Comisión Europea notificó a Google sus conclusiones preliminares, las cuales detallan las medidas propuestas para dar cumplimiento a las obligaciones de acceso a datos establecidas en el Reglamento de Mercados Digitales (DMA). Bajo este esquema, Alphabet está llamado a facilitar el acceso de motores de búsqueda de terceros —incluyendo aquellos que operan mediante *chatbots* de inteligencia artificial con funcionalidades de búsqueda— a un conjunto exhaustivo de datos sobre clasificaciones, consultas, clics y visualizaciones. Nos parece esencial señalar que, aunque en el texto de estas medidas se habla de «dataset», estas medidas conciernen esencialmente a los datos y no afectan a los *datasets* de entrenamiento utilizados por Google. Tengamos en cuenta que en este documento se denomina *dataset* de búsqueda (*Search Dataset*)⁴⁸ al producto final resultante de aplicar medidas técnicas sobre los datos originales para mitigar el riesgo de reidentificación de los usuarios

⁴⁸ Texto medidas provisionales, caso DMA.100209, pág.5. Vid. https://digital-markets-act.ec.europa.eu/document/download/b3aed7f6-c45c-4bfa-b032-b8975a48bb06_en?filename=DMA.100209_por_cien20-%20Preliminary%20measures.pdf

hasta un nivel residual. Mientras que los datos iniciales contienen identificadores y metadatos precisos, el *dataset* en este caso hace referencia a la versión procesada mediante la supresión de atributos (como IPs o marcas de tiempo exactas), la generalización de metadatos de ubicación y la aplicación de umbrales de frecuencia para garantizar la privacidad antes de su intercambio con terceros (con lo cual no encajaría en la definición de *dataset* que hemos estado manejando en este artículo).

Como vemos, en este caso el foco se centra también en el acceso a los datos procesados aunque no etiquetados.

Esta apertura se articula bajo términos justos, razonables y no discriminatorios (FRAND), con el objetivo estratégico de que estos «beneficiarios de datos» puedan optimizar sus servicios y disputar la posición de dominio de Google Search en el mercado, mediante un acceso equitativo a insumos fundamentales para la innovación tecnológica. El alcance de estas medidas, cuya efectividad está siendo sometida a una consulta pública abierta hasta el 1 de mayo, abarca tanto la determinación de la elegibilidad de los solicitantes como la definición de los medios, frecuencias y precios del intercambio de información. Asimismo, el marco regulatorio pone un énfasis especial en la seguridad jurídica y la privacidad, exigiendo la implementación de protocolos robustos de anonimización de datos personales y estableciendo procesos formales que regulen el acceso, el uso limitado a finalidades de optimización y la gobernanza de estos activos⁴⁹. Al integrar estos requisitos técnicos y contractuales, la Comisión busca asegurar que el intercambio de datos se realice de manera segura, auditable y alineada con los objetivos de contestabilidad y protección de datos que inspiran la normativa digital europea.

Más concretamente, estas medidas preliminares adoptadas por la Comisión Europea en el caso DMA.100209⁵⁰, imponen a Alphabet el deber de proporcionar sus «*Search Data*» a cualquier motor de búsqueda en línea («Online Search Engine» u OSE) operante en la Unión Europea o el Espacio Económico Europeo⁵¹. Resulta particularmente relevante, desde una perspectiva jurídica, la inclusión explícita de aquellas entidades que proveen *chatbots* de IA con funcionalidades de búsqueda; estas no pueden ser excluidas siempre que satisfagan la definición técnica de OSE, aun cuando dicha funcionalidad se despliegue como parte de un servicio más amplio.

El alcance de esta obligación se fundamenta en un estricto principio de paridad, exigiendo que Alphabet comparta los datos bajo condiciones y con

⁴⁹ https://ec.europa.eu/commission/presscorner/detail/en/ip_26_825

⁵⁰ Vid. [https://digital-markets-act.ec.europa.eu/document/download/b3aed7f6-c45c-4bfa-b032-b8975a48bb06_en?filename=DMA.100209 por cien20-%20Preliminary%20measures.pdf](https://digital-markets-act.ec.europa.eu/document/download/b3aed7f6-c45c-4bfa-b032-b8975a48bb06_en?filename=DMA.100209%20por%20cien20-%20Preliminary%20measures.pdf)

⁵¹ Esta obligación se rige por el principio de paridad, exigiendo que Alphabet comparta los datos en las mismas condiciones y con el mismo alcance con que los recopila para optimizar sus propios servicios. Los datos se clasifican en: Datos de consulta (Query data): Incluye la entrada inicial del usuario, modificaciones (autocompletado, sugerencias), metadatos de ubicación, idioma, dispositivo y punto de acceso (ej. Chrome, Gemini, Google Lens); Datos de visualización (View data): URLs y contenido visual mostrados en las páginas de resultados (SERP), incluyendo diversos formatos (video, noticias, imágenes) y módulos de búsqueda avanzada; Datos de clics e interacción (Click data): Información sobre el momento, orden y duración de las acciones del usuario, así como el tiempo de permanencia en un resultado o el regreso a la SERP; Datos de clasificación (Ranking data): La posición absoluta y relativa de cada URL en la interfaz del usuario. Texto medidas provisionales, caso DMA.100209 págs. 2-4. Vid. [https://digital-markets-act.ec.europa.eu/document/download/b3aed7f6-c45c-4bfa-b032-b8975a48bb06_en?filename=DMA.100209 por cien20-%20Preliminary%20measures.pdf](https://digital-markets-act.ec.europa.eu/document/download/b3aed7f6-c45c-4bfa-b032-b8975a48bb06_en?filename=DMA.100209%20por%20cien20-%20Preliminary%20measures.pdf)

un alcance equivalentes a los que emplea para la optimización de sus propios servicios. Este caudal informativo se categoriza en datos de consulta —que comprenden desde la entrada inicial y modificaciones por autocompletado hasta metadatos de ubicación, idioma y dispositivo—, datos de visualización, que abarcan la totalidad de URLs y contenidos mostrados en las páginas de resultados (SERP), datos de interacción o *click data*, y datos de clasificación que detallan la posición relativa y absoluta en la interfaz. La salvaguarda de la privacidad se instrumenta a través de un esquema dual: medidas técnicas de anonimización —incluyendo la supresión de identificadores directos, la generalización de metadatos mediante celdas S2 y el establecimiento de umbrales de frecuencia para evitar la reidentificación— y medidas contractuales que definen a las partes como responsables independientes del tratamiento, prohibiendo tanto la reidentificación de usuarios como el uso de datos para fines ajenos a la optimización de la búsqueda.

Las condiciones de acceso se rigen por los principios FRAND (Fair, Reasonable and Non-Discriminatory), vinculando la compensación económica a los costes incrementales y a un retorno razonable sobre el capital, con regímenes de excepción específicos para las pequeñas y medianas empresas. La eficacia de este sistema se refuerza mediante un mecanismo de auditoría independiente bajo el estándar ISAE 3000, que condiciona la obtención y mantenimiento del acceso a la presentación de informes de aseguramiento (niveles 1 y 2).

Desde nuestra perspectiva, el texto revela una vinculación directa entre estas medidas y el entrenamiento de la inteligencia artificial. Al reconocer que el propósito legítimo de estos datos incluye el anclaje de *chatbots* de IA, la normativa sitúa al «*Search Dataset*» como un insumo esencial para la competitividad en el mercado de la IA generativa aplicada a la búsqueda. Al prohibir que Alphabet margine a los desarrolladores de IA con funciones de búsqueda, la Comisión no solo promueve la contestabilidad del mercado, sino que facilita que nuevos actores refinan sus sistemas de recuperación y generación mediante el acceso a datos de interacción humana real, transformando así un recurso crítico para la innovación en un activo disponible bajo supervisión regulatoria.

3. Reflexiones en torno al caso LAION: una lectura prospectiva desde el derecho de la competencia

El caso Kneschke v. LAION⁵², que ha recibido una amplia atención desde la perspectiva del derecho de autor⁵³, ofrece una casuística especialmente rica para analizar, desde el Derecho de la competencia, la diferencia entre acceso a datos —entendidos como contenido publicado libremente en la web— y acceso a *datasets* ya estructurados, como LAION 5B, puestos a disposición de la comunidad de investigación, y cómo esta distinción influye en la delimitación de mercados, la valoración del poder de mercado y la configuración de eventuales prácticas de exclusión.

⁵² Cfr. en primera instancia, Sentencia del Landgericht Hamburg de 27 septiembre 2024, Az. 310 O 227/23 y, en segunda instancia, Sentencia del OLG Hamburg, 10.12.2025, Az. 5 U 104/24.

⁵³ Vid. entre muchos, CARBAJO CASCÓN (2026), págs. 15-52; DORNIS (2025), págs. 65-78; DORNIS y STOBER (2024), págs. 87-160; GARCÍA VIDAL (2024); LEISTNER y ANTOINE (2025), págs. 1027 - 1044, LUCCHI (2025); MANTEGUI (2025).

3.1. Contextualización del caso LAION

Como mencionábamos anteriormente, el litigio Kneschke v. LAION ha marcado un precedente crucial en la jurisprudencia alemana sobre la intersección entre el derecho de autor y el entrenamiento de modelos de inteligencia artificial (IA). El demandante, un fotógrafo profesional, alegó que la inclusión de sus obras protegidas en el *dataset* de LAION (utilizado para el entrenamiento de modelos de IA de difusión) constituía una infracción de sus derechos de propiedad intelectual, argumentando que no se cumplían los presupuestos necesarios para la aplicación de las excepciones de minería de datos (Text and Data Mining o TDM) contempladas en los arts. 44b y 60d de la Ley de Derechos de Autor alemana (UrhG), los cuales transponen la Directiva (UE) 2019/790 sobre Derechos de Autor en el Mercado Único Digital⁵⁴.

De una manera muy sucinta y centrándonos en el aspecto más esencial⁵⁵, en primera instancia el Landgericht (LG) de Hamburgo, en su sentencia de 27 de septiembre de 2024 (Az. 310 O 227/23), desestimó la pretensión del autor, al considerar la legitimidad de la aplicación del límite de TDM en el caso. Concretamente, el tribunal determinó que la actividad desarrollada por LAION, al recopilar y procesar imágenes para alimentar conjuntos de datos científicos y académicos, se subsume bajo el concepto de minería de datos legalmente permitido, incluso si los fines ulteriores de los modelos entrenados con dichos *datasets* son comerciales (Stable Diffusion era uno de los modelos que utilizaba el *dataset* de LAION).

Por su parte el Oberlandesgericht (OLG) de Hamburgo el 10 de diciembre de 2025 (Az. 5 U 104/24), ratificó esta interpretación, desestimando los argumentos del recurrente centrados en la supuesta desproporcionalidad del uso automatizado, afirmando que la normativa actual sobre TDM es suficientemente amplia para abarcar las prácticas actuales de «machine learning».

Aunque desde la perspectiva del derecho de autor, estas sentencias ofrecen un campo rico para la discusión doctrinal, el objetivo de traerlas a colación no es analizarlas desde esta perspectiva (ni de plantear la aplicabilidad o no de la excepción de TDM a la IA⁵⁶) sino de plantear una reflexión, intencionalmente provocadora, al hilo de la temática en la que se centra este artículo: los mercados de inputs y su relevancia en el actual derecho de la competencia.

⁵⁴ La Directiva (UE) 2019/790 introduce dos excepciones obligatorias al derecho de autor para facilitar la minería de textos y datos (TDM): la primera (Art. 3), de carácter imperativo, permite a los organismos de investigación y patrimonio cultural realizar TDM con fines de investigación científica sobre contenidos a los que tengan acceso lícito; la segunda (Art. 4), de carácter general, autoriza la TDM sobre obras y prestaciones accesibles legalmente, pero permite que los titulares de derechos «reserven» sus derechos para excluir tal uso —el denominado «opt-out»—, salvo que dicha reserva se realice de forma inadecuada o no sea legible por máquina. Sobre esta cuestión, ver GARCÍA VIDAL (2019), págs. 99-124; MINERO ALEJANDRE (2024), págs. 41-149; JIMÉNEZ SERRANÍA (2020), págs. 247-258.

⁵⁵ Desde luego, además de los matices de estas sentencias sobre la aplicación de la minería de textos y datos con fines analíticos del art. 3 de la Directiva (UE) 2019/790, a la generación de nueva información, tal y como hacen los modelos de IAG; también es muy interesante, entre otros aspectos, la reflexión de *ambas sentencias* sobre la manera de expresar la reserva de derechos (*opt-out*) en el caso de la aplicación de la excepción de minería de textos y datos con fines comerciales, determinándose, en segunda instancia, la necesidad de hacerlo a través de medios legibles por una máquina —por ejemplo, mediante el archivo robots.txt (y no a través del mero lenguaje humano).

⁵⁶ Sobre esta cuestión, *vid.* ROSATI (2024), págs. 851-852; LUCCHI (2025), págs. 29-72; Voss (2026), ap. K.

3.2. Reflexión final: la sustancial diferencia entre datos y dataset para el entrenamiento de IA y sus implicaciones en términos de competencia

Como hemos visto, LAION e. V. es una asociación sin ánimo de lucro que crea y distribuye *datasets* para el entrenamiento de IA—en concreto, LAION5B— a partir de datos extraídos de la web, particularmente de Common Crawl, combinando enlaces URL a imágenes disponibles públicamente y descripciones de esas imágenes (alt text, metadatos), formando así un conjunto estructurado de 5.85 mil millones de pares *imagetexto* que se utiliza para entrenar modelos de IA generativa de imágenes. Estos *datasets* no contienen las imágenes en sí, sino índices a la web, acompañados de *embeddings* calculados y puestos a disposición para la investigación científica⁵⁷.

Las sentencias de los tribunales de Hamburgo ratifican que la reproducción y tratamiento de la fotografía del demandante, en el marco de la creación de este *dataset*, queda amparada por la excepción de minería de texto y datos (TDM) del art. 3 de la Directiva DSM (2019/790) y por el correspondiente art. 60d UrhG relativo a la investigación científica, siempre que el *dataset* se ponga efectivamente a disposición de la comunidad científica sin ánimo de lucro y sin control de empresas comerciales sobre el proceso de investigación. Como decíamos, la sentencia subraya asimismo que el *optout* —la limitación de uso de contenidos— solo es eficaz si se expresa en forma máquinalegible (p. ej. robots.txt, TDMRep u otros protocolos, no meras cláusulas en lenguaje natural).

Desde la perspectiva del Derecho de la competencia, el precedente de Hamburgo permite trazar un perfil normativo útil de dos niveles:

i) Nivel de datos/obras originales: El contenido web general (imágenes, textos, vídeos) constituye un universo amplio, en principio sustituible, en el que el creador mantiene derechos de explotación, pero que el acceso técnico permite replicarse mediante *scraping* automático, siempre que se respeten las condiciones de uso y las señales técnicas de *optout*.

ii) Nivel de *dataset* LAION: El conjunto estructurado —sus índices, pares *imagetexto*, *embeddings*, y su organización en formato tabular— introduce un plus de valor agregado que no se deriva de la mera suma de obras, sino de la curación, selección, indexación y anotación aplicadas a datos de entrada masivos. Esta agregación, curación, indexación, anotación y estructuración de los datos genera un activo informacional jurídica y económicamente distinto del mero acceso al material bruto

La diferencia es, en términos competitivos, de grado de fungibilidad. Mientras que el acceso a *datos* individuales puede estar regulado por el régimen de derechos de autor y la excepción TDM, el acceso a un *dataset* estructurado como LAION5B implica un coste de entrada significativo: replicar escala, representatividad y estructura de anotación resulta extremadamente complejo y costoso, incluso si el *scraping* de datos originales se mantiene legal. LAION, por su naturaleza sin ánimo de lucro, no se sitúa como un «*player*» económico directo, pero el *dataset* que crea se convierte en insumo intermedio para mo-

⁵⁷ Vid. <https://laion.ai/faq/>

delos de IA generativa, y su estructura puede ser reutilizada por operadores comerciales.

Dicho de otra manera, la relevancia competitiva de esta diferencia no reside solo en la cantidad de datos, sino en la transformación del insumo: el *dataset* estructurado reduce costes de búsqueda, limpieza, clasificación y alineación semántica, y convierte una masa heterogénea de datos dispersos en un recurso inmediatamente utilizable para el entrenamiento. En términos de teoría del daño, el perjuicio no se agota en la eventual reproducción no autorizada de datos originales, sino que se proyecta sobre la creación de una ventaja competitiva derivada de la curaduría ajena, que puede ser apropiada por terceros sin asumir los costes fijos de construcción, mantenimiento y validación del *dataset*.

Por ello, el análisis competitivo en este tipo de casos debería formularse en términos del *dataset* como un insumo estratégico de alto valor o un recurso intermedio difícilmente replicable a igual escala y calidad. Aunque los datos brutos puedan obtenerse por vías diversas, la reproducción de un *dataset* equivalente exige costes significativos de recopilación, depuración, anotación, normalización y mantenimiento, además de recursos complementarios que no están al alcance de todos los operadores en igualdad de condiciones.

La eventual relevancia antitrust de la reutilización del *dataset* no deriva de una equiparación automática entre datos y competencia restringida, sino del efecto que su control puede tener sobre las condiciones de entrada, replicación y expansión en el mercado o en el nivel competitivo correspondiente. Si la disponibilidad de *datasets* estructurados reduce de forma sustancial los costes de desarrollo de modelos, entonces su control, exclusividad o captura de valor puede producir efectos de elevación de barreras de entrada, dependencia funcional o, en ciertos supuestos, reforzamiento de posiciones ya consolidadas.

El litigio Kneschke v. LAION muestra precisamente que el conflicto no se agota en la licitud originaria del tratamiento bajo la lógica de la TDM, sino que también puede examinarse a la luz de la posterior reutilización del *dataset* como input intermedio para el entrenamiento comercial de modelos de IA generativa. Esa reutilización puede adquirir relevancia competitiva adicional cuando permite a determinados operadores beneficiarse de una infraestructura de datos ya depurada y estructurada sin asumir los costes fijos de su construcción. No se trata, sin embargo, de calificar toda reutilización como abusiva o anticompetitiva, sino de identificar si genera una asimetría competitiva capaz de alterar las condiciones de contestabilidad del mercado.

Dicho de otro modo, la cuestión no es únicamente si existe acceso a datos, sino qué tipo de activo informacional se está controlando. Los datos brutos y los *datasets* de entrenamiento para IA presentan funciones económicas distintas y, por ello, pueden producir efectos competitivos diferenciados. Mientras los primeros integran materiales informativos potencialmente reutilizables, los segundos condensan una capa adicional de selección, organización y procesabilidad que puede modificar de forma relevante las condiciones de acceso, replicación y uso dentro de la cadena de valor de la IA. Por ello, su relevancia competitiva no depende de su equiparación al dato bruto, sino de su función como recurso intermedio estructurado, capaz de generar ventajas de entrada, dependencia o exclusión en condiciones determinadas. En consecuencia, el análisis antitrust,

en nuestra opinión, debe contemplar como elementos diferenciados los *datasets* de entrenamiento de IA dentro de la cadena de valor.

V. BIBLIOGRAFÍA

- ACEMOGLU, Daron y RESTREPO, Pascual (2019) «Automation and New Tasks: How Technology Displaces and Reinstates Labor», *Journal of Economic Perspectives* 33 (2), págs. 3–30.
- AKMAN, Pinar, FUMAGALLI, Chiara y MOTTA, Massimo, «The European Commission's draft guidelines on exclusionary abuses: a law and economics critique and recommendations», *Journal of European Competition Law & Practice*, Vol. 16, I. 4, págs. 234–243.
- ALAILY, Rima (2024), «The New AI Economy: Understanding the Technology, Competition, and Impact for Societal Good», *Concurrences*, N° 2-2024 – AI & Antitrust, págs. 5-10.
- CARBAJO CASCÓN, Fernando, (2026), «Sobre la aplicación de los límites de minería de textos y datos al proceso de entrenamiento de modelos de inteligencia artificial generativa de uso general» en LÓPEZ ORTEGA, Raquel y PACHECO CAÑETE, Matilde (dir.), *Nuevos horizontes del Derecho mercantil en un entorno tecnológico*, Atelier, Barcelona, págs. 15-52.
- CARUGATI, Ch. (2023A), *The competitive relationship between cloud computing and generative AI*, Working Paper, Bruegel, Bruselas.
- (2023B), *Competition in Generative AI Foundation Models*”, Working Paper 14/2023, Bruegel, Bruselas.
- COLANGELO, Giuseppe (2025) «The EU Essential Facilities Doctrine after Android Auto: A Wild Card without Limiting Principles?», *CMLRev*, Vol.62, I. 5, págs. 1449-1476.
- DANG, Andrew, (2025), «Derivative Data: Rethinking Market Definitions in the Age of Generative AI», 16 *UC LAW SF SCI. & TECH. JUST. L.J.* 97, págs. 120-127.
- DORNIS, Tim W. y STÖBER, Sebastian (2024), *Urheberrecht und Training generativer KI-Modelle, Technologische und juristische Grundlagen*, Nomos: Recht und Digitalisierung; Digitization and the Law, Vol. 19.
- DORNIS, Tim, (2025) «The training of Generative AI is not Text and Data Mining», *EIPR*, 2/2025, vol. 47(2), págs. 65-78.
- DREXL, Josef y KIM, Daria (2026), «AI Innovation Competition as a Discovery Procedure: The Role and Limits of Competition Law» en Arezzo, Emanuela (ed), *Navigating the (Legal) Challenges of the Artificial Intelligence Era: Intellectual Property, Competition Law and Corporate Law*, págs. 181-218.
- EDELSON, Laura, GRAEF, Inge y LANCIERI, Filippo (2023), *Access to Data and Algorithms: For an Effective DMA and DSA Implementation*, CERRE, <https://cerre.eu/publications/access-to-data-and-algorithms-for-an-effective-dma-and-dsa-implementation>.
- GÁLVEZ DURAN, Sergi (2025), «Mapping relevant data collection mechanisms for AI training», *OECD Artificial Intelligence Papers*, Octubre 2025, n°48. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/10/mapping-relevant-data-collection-mechanisms-for-ai-training_62921889/3264cd4c-en.pdf.
- GARCÍA VIDAL, Ángel (2019), «Propiedad intelectual y minería de textos y datos. Estudio de los artículos 3 y 4 de la Directiva (UE) 2019/790», *ADI*, Vol. 40, págs. 99-124.
- (2024), *Propiedad intelectual, minería de textos y datos y entrenamiento de la inteligencia artificial*, accesible en https://ga-p.com/wp-content/uploads/2024/10/Mineria_textos_datos.pdf.
- HUGGINS, Robert, JOHNSTON, Andrew, MUNDAY, Max y XU, Chen (2023) «Competition, open innovation, and growth challenges in the semiconductor industry: the

- case of Europe's clusters», *Science and Public Policy*, Vol. 50, I. 3, págs. 531-547.
- JIMÉNEZ SERRANÍA, Vanessa (2020), «Datos, minería e innovación: qvo vadis Europa», *CDT*, Vol. 12, núm. 1, págs. 247-258.
- (2026), «La definición del mercado relevante en la era de la inteligencia artificial: ¿un (no tan) nuevo reto para el derecho de la competencia?» en TATO PLAZA, Anxo, COSTAS COMESAÑA, Julio, FERNÁNDEZ CARBALLO-CALERO, Pablo, TORRES PÉREZ, Francisco José y LOUREDO CASADO, Sara (dirs.), *Nuevas tendencias en el Derecho de la competencia y de la Propiedad Industrial e Intelectual V*, Atelier, Barcelona, págs. 145-172.
- KERBER, Wolfgang (2023), *Towards a dynamic concept of competition that includes innovation*, OCDE, DAF/COMP/WD(2023) 42, págs. 12-15.
- KOWALSKI, Klaus, VOLPIN, Cristina y ZOMBORI, Zsolt (2024), *Competition in Generative AI and Virtual Worlds*, Competition Policy Brief, European Union, https://competition-policy.ec.europa.eu/document/download/c86d461f-062e-4dde-a662-15228d6ca385_en.
- LEISTNER, Matthias y ANTOINE, Lucie (2025), «TDM and AI training in the European Union – from »LAION” to possible ways ahead?”, *GRUR International*, Vol. 74, I 11, págs. 1027 - 1044.
- LIU, Z Zhifeng, LUO, Pengcheng, TANG, Xinglong *et al.* (2024), «Unfolding the downloads of datasets: A multifaceted exploration of influencing factors», *Sci Data* 11, 760.
- LUCCHI, Nicola (2025), *Generative AI and Copyright (training, creation, regulation)*, accesible en [https://www.europarl.europa.eu/thinktank/en/document/IUST_STU\(2025\)774095](https://www.europarl.europa.eu/thinktank/en/document/IUST_STU(2025)774095);
- MANTEGUI, Maryna (2025), «Kneschke vs. Laion: Opening Fresh Perspectives on AI Training and TDM Exceptions», accesible en <https://www.europeanlawblog.eu/pub/tk4iacho/release/1>.
- MINERO ALEJANDRE, Gemma (2024), «Los dos límites (el obligatorio y el optativo) de minería de textos y datos», en CÁMARA ÁGUILA, Pilar y GARROTE FERNÁNDEZ-DÍEZ, Ignacio (Dirs), *La transposición en España de la Directiva sobre derechos de autor y mercado único digital*, Tirant lo Blanch, Valencia, págs. 41-149.
- PEUKERT, Christian (2025), *The economics of copyright and AI. Empirical evidence and optimal policy*, European Union, págs.17 y 18.
- PICHT, Peter G. (2025), Data licensing and data-related provisions in technology transfer agreements, European Union, https://competition-policy.ec.europa.eu/document/download/9eff6170-627c-4d31-81f8-9d2317f41ea0_en?filename=kd0125011enn_TTBER_study.pdf
- RADIC, Lazar y STOUT, Kristian (2024), «What is the Relevant Product Market in AI?» en ABBOTT, Adel y SCHREPEL, Thibault. (eds.), *Artificial Intelligence and Competition Policy*, Concurrences, Paris, págs. 107-132.
- RIBERA MARTÍNEZ, Alba (2025), «The Regulation that Cried Wolf: Generative AI Training Data and the Challenge of Lawful Scale» *Forthcoming in Journal of European Competition Law & Practice*. Accesible en https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5222181.
- ROSATI, Eleonora (2024), «Is text and data mining synonymous with AI training?», *JIPLP*, Vol. 19, No. 12, págs. 851-852.
- SCHREPEL, Thibault y PENTLAND, Alex «Sandy» (2025), «Competition between AI foundation models: dynamics and policy recommendations», *Industrial and Corporate Change*, Vol. 34, I. 5, págs. 1086 y 1087.
- STUCKE, Maurice E. y EZRACHI, Ariel (2025) “Antitrust & AI supply chains”, *Theoretical Inquiries in Law*, vol. 26, no. 1, págs. 147-182.
- THIEMANN, Ania, Artificial intelligence and competitive dynamics in downstream markets, OCDE, DAF/COMP/GF(2025)4, Paris, <https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/11/artificial-intelligence-and-competitive-dyna>

- mics-in-downstream-markets_c6e81d0e/ccf0624a-en.pdf*;
- VOSS, Axel (2026) *Report on copyright and generative artificial intelligence- opportunities and challenges*, Parlamento Europeo, A10-0019/2026, https://www.europarl.europa.eu/doceo/document/A-10-2026-0019_EN.html.
- ZIERMANN, Fabian (2025), «Defining data markets: an AI-centric gaming analogy rooted in market realities», *Journal of European Competition Law & Practice*, págs. 352–361.