3.
Innovación pública y Administración digital

El problema de la interpretabilidad de la Inteligencia Artificial y su impacto en la Administración Pública

Silvia Alayón Miranda
Profesora Titular del área de Ingeniería de Sistemas y Automática.
Departamento de Ingeniería Informática y Sistemas.
Universidad de La Laguna

RESUMEN: La Inteligencia Artificial (IA) es una de las tecnologías más disruptivas y transformadoras de la actualidad. Está cada vez más presente en todos los ámbitos, y este uso masivo de la tecnología nos está cambiando: está cambiando nuestra formación, nuestro modo de trabajar, y nuestra forma de relacionarnos entre nosotros y con nuestro entorno. Por supuesto, la Administración Pública no puede (ni debe) ser ajena a todos estos cambios.

En un futuro no muy lejano, gran parte de las decisiones que afectan a los ciudadanos estarán tomadas por algoritmos inteligentes. En este escenario la interpretabilidad de los algoritmos de IA será imprescindible para garantizar su buen uso en la Administración Pública.

Este artículo pretende explicar qué es la interpretabilidad de la IA y por qué es importante, cómo se está abordando esta cuestión desde el punto de vista técnico, cuál es el papel de la regulación en la interpretabilidad de la IA y cuáles son las perspectivas de futuro, sobre todo en el ámbito de la Administración Pública. Para dar un enfoque lo más completo posible, se explicarán algunos fundamentos básicos de la tecnología involucrada y del grado de integración actual de la IA en Europa y en nuestro país.

Palabras clave: Inteligencia Artificial, Administración Pública, Interpretabilidad, Explicabilidad, Transparencia.

ABSTRACT: Artificial Intelligence (AI) is one of today's most disruptive and transformative technologies. It is increasingly present in all areas, and this massive use of technology is changing us: it is changing our education, the way we work, and the way we relate to each other and to our environment. Of course, Public Administration cannot (and should not) be oblivious to all these changes.

In the not too distant future, a large part of the decisions affecting citizens will be made by intelligent algorithms. In this scenario, the interpretability of AI algorithms will be essential to ensure their proper use in public administration.

This article aims to explain what AI interpretability is and why it is important, how this issue is being addressed from a technical point of view, what is the role of regulation

in AI interpretability and what are the future prospects, especially in the field of Public Administration. In order to offer the most comprehensive approach possible, some basic fundamentals of the technology involved and the current degree of integration of AI in Europe and in our country will be explained.

Keywords: Artificial Intelligence, Public Administration, Interpretability, Explainability, Transparency.

SUMARIO: 1. INTRODUCCIÓN. 2. LA ADMINISTRACIÓN PÚBLICA EN ESPAÑA Y LA TRANSFORMACIÓN DIGITAL. 3 TECNOLOGÍAS INVOLUCRADAS EN LAS SOLUCIONES TECNOLÓGICAS PARA LA ADMINISTRACIÓN PÚBLICA. 4. EL PROBLEMA DE LA INTERPRETABILIDAD 4.1. Definición. 4.2. Tipos de interpretabilidad y tipos de técnicas de interpretación. 4.3. Un ejemplo claro: el problema de la interpretabilidad en las aplicaciones de IA diseñadas para el entorno médico. 4.4. Problemas de interpretabilidad en aplicaciones de IA diseñadas para la Administración Pública. 5. LA CONFIANZA EN LA IA POR PARTE DE LOS USUARIOS. 6. APLICACIONES DE IA ACTUALMENTE ACTIVAS EN LA ADMINISTRACIÓN PÚBLICA ESPAÑOLA. CONCLUSIONES. BIBLIOGRAFÍA

1. INTRODUCCIÓN

Empecemos por el principio… ¿qué es exactamente la Inteligencia Artificial (IA)? Según la RAE, la IA es una *"disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico"*. Es decir, una máquina o programa de IA es capaz de aprender, tomar decisiones y realizar tareas de manera autónoma, tal y como hacemos los humanos. Este objetivo, planteado en los años 50 del siglo pasado (Moor, 2006), es, en mi opinión, la meta más ambiciosa de la Informática. Una meta que parece ya casi alcanzada con los grandes avances de estos últimos 4-5 años.

El gran potencial que tienen los programas de IA para aprender a realizar tareas como las personas en múltiples contextos ha disparado su aplicación en todos los campos del conocimiento. Con el fin de hacernos la vida más fácil, la IA se aplica ya en innumerables ocasiones para automatizar tareas repetitivas y temporalmente costosas. Este tipo de sistemas, diseñados para realizar tareas específicas y limitadas, replicando el comportamiento humano sólo en ese contexto y de manera restringida, son ejemplos de **IA débil** (Rouse, 2024a). En cambio, la **IA fuerte** (Rouse, 2024b), que es el tipo de IA que está en pleno desarrollo en estos momentos, persigue replicar la inteligencia humana sin restricciones. Las aplicaciones de IA fuerte son capaces de aprender de nuevas situaciones, razonar e inferir nuevos conocimientos, y tomar decisiones, tal y como lo hacemos los humanos. Ejemplos de esto son las últimas aplicaciones de IA en Medicina, capaces de realizar diagnóstico de manera autónoma (Kim et al., 2022; Kumar et al., 2023) e incluso predicciones sobre la evolución de la enfermedad (Yala et al., 2019; Zhou et al., 2023), o las aplicaciones de IA generativas, que son capaces

de crear nuevos contenidos (texto, imágenes, videos, etc.) (Gallent et al., 2023; Franganillo, 2023).

Centrándonos en la aplicabilidad de la IA en la Administración Pública, la distinción entre IA débil o IA fuerte es de especial relevancia. Por una parte, parece evidente el gran número de ventajas que una aplicación de IA débil puede ofrecer al desempeño diario de los funcionarios de cualquier órgano: en todos los ámbitos de la Administración Pública hay un sinfín de procedimientos rutinarios ya establecidos por protocolos que consumen mucho tiempo y recursos de personal que podrían ser realizados por este tipo de aplicaciones, liberando así a las personas para que se pudieran ocupar de otras tareas más complejas e importantes. En cambio, cuando analizamos la posibilidad de integrar alguna herramienta de IA fuerte en la actividad de cualquier administración, el debate es más complicado. Surgen dudas sobre si una aplicación de IA que parece capaz de tomar decisiones es realmente fiable, cuando esas decisiones afectan directamente a los usuarios de un servicio público: ¿será capaz de tomar decisiones coherentes y equitativas para todos los usuarios de ese servicio? ¿será entrenada con datos que garanticen un comportamiento ético? Y si la aplicación de IA se equivoca… ¿quién es el responsable?

Todas estas dudas son consecuencia de la falta de transparencia de los modelos de IA, que se utilizan en la práctica como si fueran "cajas negras". Sabemos que funcionan muy bien en un alto porcentaje de situaciones, pero no entendemos por qué, no somos capaces de desentrañar el interior del modelo para poder comprender, interpretar o explicar los motivos que le han llevado a tomar una decisión u otra. Por este motivo, cuando nos planteamos introducir aplicaciones de IA fuerte en ámbitos delicados, surgen las dudas relativas a la fiabilidad del modelo: ¿podemos confiar en él si no entendemos cómo funciona?

Actualmente, la investigación en IA avanza a pasos agigantados: cada vez hay modelos más precisos para aplicaciones muy diversas. Pero es un hecho que en los problemas de aprendizaje automático **los modelos más precisos son los más complejos**. Esto es así porque precisamente estos modelos buscan patrones que captan relaciones entre muchas variables en espacios de alta dimensionalidad y, por tanto, son capaces de "ver" relaciones e información muy difícil de captar e interpretar por un ser humano. Esta es su principal fortaleza, pero también su mayor debilidad: los mejores modelos son normalmente los menos interpretables. Este problema, conocido como **el problema de la interpretabilidad** de la IA (Linardatos et al., 2020), es actualmente un frente abierto en la investigación, y uno de los mayores obstáculos que impide una mayor integración de la IA en la Administración Pública (y en muchos otros ámbitos).

Este artículo pretende explicar qué es la interpretabilidad de la IA y por qué es importante, cómo se está abordando esta cuestión desde el punto de vista técnico, cuál es el papel de la regulación en la interpretabilidad de la IA y cuáles son las perspectivas de futuro, sobre todo en el ámbito de la Administración Pública. Para dar un enfoque lo más completo posible, se explicarán algunos fundamen-

tos básicos de la tecnología involucrada y del grado de integración actual de la IA en Europa y en nuestro país.

2. LA ADMINISTRACIÓN PÚBLICA EN ESPAÑA Y LA TRANSFORMACIÓN DIGITAL

La automatización de los procesos realizados en la Administración Pública es un tema que empezó a despertar interés en la comunidad investigadora hace aproximadamente 20 años, cuando se constató la necesidad real de prestar al ciudadano un servicio ágil y personalizado, intentando reducir los tiempos de espera y aumentando la información ofrecida del proceso administrativo. Por este motivo, surgieron en España los primeros trabajos que plantearon estrategias para automatizar el ciclo de vida del expediente (Mariano, 2002), y se diseñaron los primeros "Sistemas de Gestión Integral de Expedientes (SGIE)" (Rodríguez y González, 2002).

Con el auge de las Tecnologías de la Información y las Comunicaciones (TIC), este interés inicial en la automatización de procesos administrativos se fue consolidando hasta llegar a englobarse en la actual "Transformación Digital" en la que estamos inmersos. Es evidente que tecnologías emergentes como el Internet de las Cosas, la Inteligencia Artificial, el Cloud Computing, etc. están cambiando todos los aspectos de la sociedad, y la gestión de los procesos en la Administración Pública no se debe quedar atrás.

La Trasformación Digital es tan importante que es la protagonista de uno de los objetivos prioritarios de la Comisión Europea para el periodo 2019-2024: hacer de la década del 2020 la "Década Digital Europea", y consolidar la soberanía digital de Europa estableciendo sus propias normas, centradas particularmente en los datos, la tecnología y las infraestructuras (Comisión Europea, 2019). Con respecto a los servicios públicos, el objetivo de la Comisión es conseguir que, para 2030, todos los servicios públicos clave sean accesibles en línea, que los ciudadanos puedan tener acceso a su historial médico electrónico y que el 80% de estos ciudadanos ya estén utilizando una solución de identificación electrónica.

El Gobierno de España no es ajeno a esta situación, y por este motivo ha impulsado "España Digital 2026" (Gobierno de España, 2024a), la actualización de la estrategia lanzada en julio de 2020 como hoja de ruta de la transformación digital del país. La estrategia aborda diversos aspectos, y, dentro del apartado "Economía", destaca el punto "5. Transformación digital del sector público: Impulsar la digitalización de las Administraciones Públicas, particularmente en ámbitos clave como el Empleo, la Justicia, o las Políticas Sociales, mediante la actualización de las infraestructuras tecnológicas".

Es importante destacar que este interés del Gobierno de España en automatizar tareas en la Administración es anterior a la publicación de esta agenda. De hecho, la Ley 11/2007, de 22 de junio ya regula el acceso electrónico de los

ciudadanos a los servicios públicos. Desde entonces, la Administración ha ido incorporando nuevas soluciones tecnológicas en su funcionamiento, tal y como se puede consultar en el Catálogo de servicios de Administración digital de la a Secretaría General de Administración Digital (SGAD) (Gobierno de España, 2024b). Las soluciones tecnológicas actualmente ofrecidas se agrupan en diferentes categorías: "Impulso de la Administración Digital y de Servicios a los Ciudadanos", "Gestión interna", "Infraestructuras", y "Regulación, guías e informes relevantes".

Por su parte, el Gobierno de Canarias, a través de la Consejería de Presidencia, Justicia e Igualdad, suscribió el 14 de julio de 2016 un convenio de colaboración con la Administración General del Estado para la prestación mutua de soluciones básicas de administración electrónica (publicado en BOE nº 217 de 8 de septiembre de 2016) donde se establecieron las condiciones en las que la Administración Pública de la Comunidad Autónoma de Canarias facilita el acceso a las Entidades Locales y a las entidades de derecho público vinculadas o dependiente de las mismas a los siguientes servicios: sistemas de firma electrónica avanzada, comunicaciones entre Administraciones Públicas por medios electrónicos, notificación por medios electrónicos, etc.

Todas estas soluciones, en línea con los avances impulsados desde Europa, hacen uso de la tecnología para agilizar y automatizar muchas tareas comunes de la actividad de la Administración. Algunas de estas aplicaciones implican simplemente la posibilidad de intercambiar información electrónicamente entre agentes, y otras, involucran IA débil, ya que el uso de la IA en este contexto se desarrolla de manera controlada, para automatizar algunas tareas rutinarias. El uso de aplicaciones de IA fuerte, donde una máquina pueda llegar a tomar decisiones de gran responsabilidad, aún no está normalizado en la Administración, a pesar de que existen propuestas técnicas (Veale y Brass, 2019; Restrepo-Amariles, 2020; Anastasopoulos y Whitford, 2019; Henman, 2020). Hay dudas importantes relativas a la interpretabilidad de este tipo de sistemas, cuestiones éticas y lagunas legales en lo que atañe a la responsabilidad de las decisiones tomadas por estos sistemas (Cerrillo i Martínez, 2019).

3. TECNOLOGÍAS INVOLUCRADAS EN LAS SOLUCIONES TECNOLÓGICAS PARA LA ADMINISTRACIÓN PÚBLICA

El constante aumento de las posibilidades técnicas hace que la automatización de procesos resulte cada vez más atractiva para la Administración Pública. Gracias a los avances de la Inteligencia Artificial (IA), hoy se pueden automatizar procesos que hace solo unos años tenían que ser realizados por humanos. Pero no todos los procesos administrativos pueden automatizarse del mismo modo desde un punto de vista técnico (Etscheid, 2019; Sobrino-García, 2021). Es necesario estudiar qué tecnologías pueden llegar a ser parte de la solución óptima para cada caso, puesto que la IA es muy amplia, y se compone de muchas

subdisciplinas. A continuación, describiremos brevemente las más aplicadas en el contexto analizado.

Una herramienta fundamental para la implementación de automatizaciones inteligentes es la matriz de asignación de responsabilidades o matriz RACI (*Responsible, Accountable, Consulted, Informed*). Las matrices RACI se concibieron específicamente para indicar el nivel de responsabilidad que tiene cada recurso humano con respecto a cada actividad realizada en una organización, desde el ejecutor del trabajo hasta el recurso que debe aprobarlo o recibir determinadas notificaciones (Cabanillas et al., 2012). En el ámbito de la Administración, la matriz RACI puede facilitar la identificación de quiénes son los responsables de las distintas actividades del servicio, y de las decisiones relacionadas con la gestión y tramitación de expedientes y/o movimiento de otros datos dentro del mismo.

Una vez identificados los datos y los agentes involucrados en el proceso administrativo, ya es posible plantear el uso de técnicas más avanzadas para el modelado y la automatización del mismo. La tecnología más disruptiva en estos momentos es la **Automatización Robótica de Procesos** (**RPA** – *Robotic Process Automation*), que es una tecnología emergente de automatización de los procesos de una empresa, organización, etc. que replica las acciones que un ser humano realiza durante dicho proceso (Mullakara y Asokan, 2020). En el contexto de las actividades actuales hacia la modernización administrativa basada en la digitalización de los procesos, el uso y la integración de software de RPA en los procesos de trabajo de la Administración Pública puede mejorar significativamente la eficiencia, reducir los costes de los procesos y mejorar el servicio dado al ciudadano. Algunos ejemplos recientes de su aplicación se pueden consultar en (Mullakara y Asokan, 2020; Houy et al., 2019; Uskenbayeva et al., 2019; Johansson et al., 2022). En este último trabajo, además, se plantea que las herramientas RPA pueden incluso llegar a ser una guía de la buena burocracia a seguir dentro de la organización.

Dependiendo de cómo sea la naturaleza de los datos involucrados en el proceso, o de las tareas que implique el mismo, puede ser necesario el uso adicional de otras tecnologías complementarias. Por ejemplo, el **Procesamiento de Lenguaje Natural** (**NLP** – *Natural Languaje Processing*) está ganando terreno en la investigación sobre gestión y automatización de procesos por su capacidad para analizar y comprender automáticamente el lenguaje humano (Kang et al., 2020; Kowalski et al., 2017). Es tal su potencial impacto que la Unión Europea ha generado una guía de directrices y buenas prácticas específica para las organizaciones públicas que estén interesadas en iniciar un proyecto de NLP (Comisión Europea, 2022).

Por otra parte, las técnicas de procesamiento de macrodatos o **Big Data** podrían también tener cabida en este tipo de soluciones tecnológicas. Esta tecnología, capaz de recopilar y analizar grandes volúmenes de datos y reconocer patrones y tendencias en ellos, ha atraído enormemente la atención de los inves-

tigadores en ciencias de la información, los responsables políticos y los responsables de la toma de decisiones en organizaciones públicas y empresas privadas (Chen, 2016). Los Big Data son conjuntos de datos tan grandes y complejos que precisan de aplicaciones informáticas no tradicionales de procesamiento para tratarlos adecuadamente, como, por ejemplo, las técnicas de **Aprendizaje Automático** (**ML** – *Machine Learning*).

En los últimos años, el **Aprendizaje Profundo** (**DL** – *Deep Learning*), englobado dentro del Aprendizaje Automático, ha permitido realizar tareas de reconocimiento de patrones, de imágenes, de voz y procesamiento de vídeo de manera automática con un rendimiento espectacular (Sarker, 2021). Los modelos DL se basan en arquitecturas de redes neuronales artificiales. Una red neuronal artificial consta de unidades de cómputo (neuronas) distribuidas en capas e interconectadas entre sí. La señal de entrada se va propagando desde la capa de entrada hasta la de salida, siendo procesada por las neuronas de las capas intermedias (capas ocultas). El número de capas ocultas puede ser variable. Cuantas más capas tiene la red, más profunda (y, por tanto, más compleja) es, de ahí el término "deep" que caracteriza estas redes (Torres, 2020). Los modelos de Deep Learning pueden tener cientos o incluso miles de capas ocultas (figura 1).

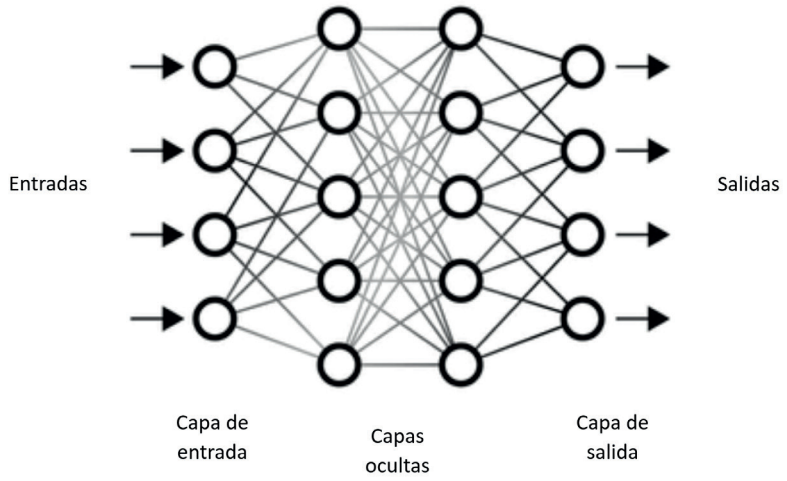

Figura 1. Esquema básico de la arquitectura de una red DL.

Aunque la primera red neuronal, denominada *perceptrón*, surgió en 1958, y en años posteriores se desarrolló la teoría relativa a estas redes neuronales artificiales, no fue hasta la década del año 2000 cuando se comenzó a aprovechar su verdadero potencial. Hasta ese momento, los recursos informáticos eran insuficientes, además de existir una gran dificultad para conseguir los datos requeridos en los entrenamientos de estos sistemas. La aparición de las GPU de alto rendimiento, las posibilidades de cálculo en la nube, el gran desarrollo de

Internet, la facilidad para compartir datos y la cultura de acceso abierto que actualmente impera en el campo de la IA ha hecho que estas redes crezcan de manera explosiva en los últimos 5 años.

Algunos organismos y agencias públicas han aprovechado estas técnicas para intentar automatizar algunos servicios públicos (Kowalski et al., 2017). Más recientemente, sin embargo, ha surgido la idea de aplicar estos modelos de Aprendizaje Profundo sobre los datos administrativos, para construir modelos que ayuden a tomar decisiones operativas cotidianas en la gestión y prestación de los servicios públicos (Veale y Brass, 2019; Restrepo-Amariles, 2020; Anastasopoulos y Whitford, 2019; Henman, 2020).

Todas estas tecnologías tienen algo en común: **necesitan datos**. Las aplicaciones de IA son capaces de imitar el comportamiento humano cuando son entrenadas con un número alto de datos descriptivos del problema abordado. El número de datos utilizado en el entrenamiento es un factor muy importante, así como la calidad de estos: deben ser variados y representativos de todas las situaciones que pueden ocurrir dentro del proceso a automatizar (Felderer y Ramler, 2021). Sin embargo, la recopilación de estos datos puede ser difícil en el contexto que nos ocupa: en muchas situaciones los procesos de la Administración Pública manejan información privada de usuarios, que no puede ser difundida o utilizada públicamente, protegida por la Ley de protección de datos (Martens, 2018). La segunda dificultad es garantizar la calidad de los datos: es importante asegurarse de que los datos sean precisos, completos y relevantes para la automatización del proceso estudiado. En este sentido, el Ministerio de Asuntos Económicos y Transformación Digital ha publicado algunas recomendaciones en su Guía al Análisis Exploratorio de Datos (Gobierno de España, 2021). El Análisis Exploratorio de Datos (AED) consiste en aplicar un conjunto de técnicas estadísticas dirigidas a explorar, describir y resumir la naturaleza de los datos, de tal forma que se pueda garantizar su objetividad e interoperabilidad.

4. EL PROBLEMA DE LA INTERPRETABILIDAD

4.1. Definición

La "interpretabilidad", en el contexto de IA, es la capacidad de **explicar** el comportamiento de los sistemas de IA de manera comprensible para un humano (Doshi-Velez y Kim, 2017). Es lo que se conoce también como Inteligencia Artificial Explicable (en inglés: *Explainable Artificial Intelligence*, abreviado XAI) (Saeed y Omlin, 2023). De ahí que en la literatura científica se usen indistintamente los términos "interpretabilidad" y "explicabilidad". Otro modo de indicar que un sistema de IA es interpretable o explicable, es decir que es "transparente", para destacar que es lo opuesto a un sistema de IA con el que se trabaja como si fuera una "caja negra" (un sistema de IA que ante determinadas entradas ofrece la salida deseada, pero sin saber el usuario cómo lo consigue).

Históricamente, la necesidad de explicaciones en el área de la IA se remonta a los primeros trabajos de explicación de sistemas expertos (Biran y Cotton, 2017). Sin embargo, ha sido en los últimos años, con el boom de los sistemas de IA basados en Aprendizaje Profundo (Deep Learning, DL), cuando la XAI ha ganado relevancia para los investigadores.

A medida que los modelos de IA han ido mejorando, se ha transitado de la IA débil (diseñada para realizar tareas específicas y limitadas, replicando el comportamiento humano sólo en ese contexto y de manera restringida) a la IA fuerte (capaz de replicar la inteligencia humana sin restricciones: aprender de nuevas situaciones, razonar e inferir nuevos conocimientos, tomar decisiones, crear, etc.). Estas aplicaciones de IA fuerte son las que tienen el potencial de revolucionar muchos ámbitos relevantes en nuestra vida, y por este motivo, son las que han disparado la alarma de la interpretabilidad (figura 2).

Los humanos necesitamos comprender cómo funciona un mecanismo para poder confiar él, sobre todo si de ese mecanismo pudiera depender nuestra vida. Por este motivo, el no entender cómo funcionan estos sistemas IA avanzados es lo que está frenando su integración en entornos tan sensibles como son la Medicina, la Educación, la Administración Pública, el Transporte, etc, además de despertar miedo en los expertos por los riesgos que conlleva un uso descontrolado de este tipo de sistemas (Gladstone AI, 2024).

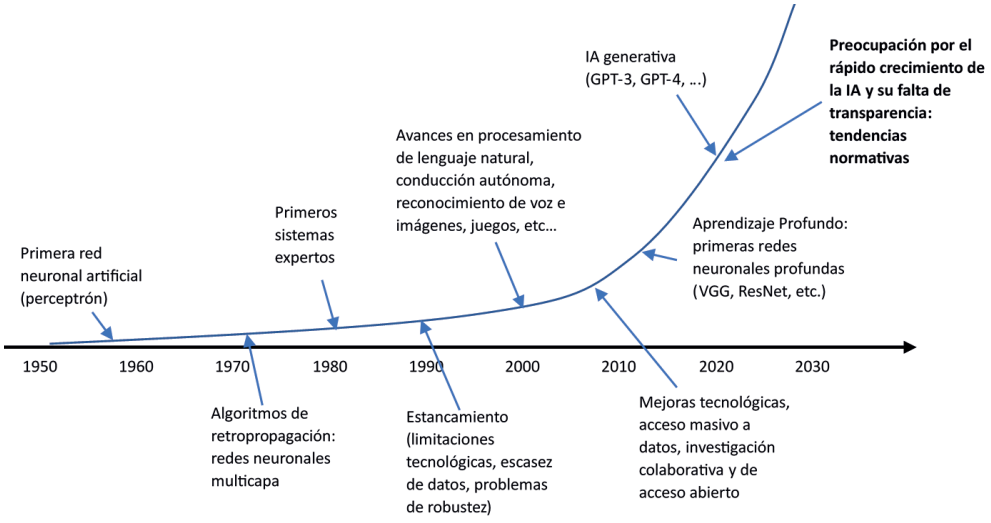

Figura 2. Evolución de la Inteligencia Artificial.

El problema de la interpretabilidad de estos sistemas es tan grave, que hay científicos, encabezados por Elon Musk, que han pedido un alto en la investigación y desarrollo de sistemas IA para intentar entender su funcionamiento interno y así, poder controlarlo adecuadamente:

«*Los laboratorios de IA y los expertos independientes deberían aprovechar esta pausa para desarrollar y aplicar conjuntamente una serie de protocolos de seguridad compartidos para el diseño y desarrollo de la IA avanzada, rigurosamente auditados y supervisados por expertos externos independientes. Estos protocolos deben garantizar que los sistemas que se adhieran a ellos sean seguros más allá de toda duda razonable. Esto no significa una pausa en el desarrollo de la IA en general, simplemente* ***un paso atrás en la peligrosa carrera hacia modelos de caja negra impredecibles cada vez más grandes y con capacidades emergentes***» (Woollacott, 2023).

En su carta piden específicamente una pausa de al menos seis meses en el entrenamiento de sistemas de IA más potentes que el GPT-4 de Open AI, con los gobiernos preparados para intervenir y aplicar una prohibición.

4.2. Tipos de interpretabilidad y tipos de técnicas de interpretación

Cuando nos referimos a la interpretabilidad de un modelo IA, se pueden diferenciar dos tipos de interpretabilidad: la **interpretabilidad global** y la **interpretabilidad local** (Du et al., 2019):

- La interpretabilidad global se obtiene cuando el usuario es capaz de comprender cómo funciona el modelo a nivel global inspeccionando las estructuras y parámetros internos del mismo.
- La interpretabilidad local analiza una predicción individual del modelo, e intenta explicar qué característica de entrada dio lugar a la decisión particular de esa predicción.

En el caso de los modelos basados en Aprendizaje Profundo (DL - *Deep Learning*), que son los que están actualmente en plena expansión, la interpretabilidad global se alcanza si se pueden comprender las representaciones capturadas por las neuronas en cualquier capa intermedia del modelo (Yosinski et al., 2015; Nguyen et al., 2016), mientras que la interpretabilidad local se obtiene al identificar las contribuciones de cada característica de una entrada concreta en la predicción realizada por el modelo (Ancona et al., 2018; Selvaraju et al., 2017; Wang et al., 2019).

En general, las técnicas de interpretación de los modelos IA se dividen en dos categorías: **técnicas intrínsecas** y **técnicas post-hoc** (Du et al., 2019), dependiendo del momento temporal en el que se obtiene la interpretación del modelo:

- La interpretabilidad intrínseca se obtiene cuando es posible construir modelos auto-explicativos (*self-explanatory models*), modelos que incorporan en su estructura elementos interpretables.
- Las técnicas *post-hoc* requieren el diseño de un segundo modelo para poder analizar y explicar el primer modelo.

La principal diferencia entre estas dos aproximaciones es la relación entre la eficiencia del modelo (*accuracy model*) y la fidelidad de la explicación (*explanation fidelity*). En las técnicas intrínsecas la calidad de la explicación es mejor, pero se sue-

le sacrificar la eficiencia de predicción del modelo, ya que al introducir elementos explicativos en el mismo se altera su estructura interna. En las técnicas post-hoc no se altera el modelo original, por lo que se puede conseguir su máxima eficiencia, pero las explicaciones obtenidas a posteriori del mismo son menos transparentes.

Con respecto a los sistemas de Aprendizaje Profundo (DL), la mayor parte de las técnicas existentes son post-hoc (Yosinski et al., 2015; Nguyen et al., 2016; Ancona et al., 2018; Selvaraju et al., 2017; Wang et al., 2019), por dos motivos principales: en primer lugar, es difícil introducir elementos explicativos en las estructuras internas del modelo, y en segundo lugar, la principal ventaja de estos modelos es su alta eficiencia, no interesa poder comprender el comportamiento interno de un modelo peor, interesa interpretar el modo en el que el mejor modelo ha conseguido ese valor tan alto de precisión. Considerando todo esto, comentaremos a continuación las técnicas de explicabilidad post-hoc más utilizadas.

Los métodos de perturbación funcionan perturbando la entrada y viendo qué efecto tiene dicha perturbación en la decisión final del modelo de IA (Zeiler y Fergus, 2013; Bach et al., 2015; Samek et al., 2015). Estas técnicas intentan descubrir qué características de una imagen de entrada son las más influyentes en la decisión del modelo. El gran inconveniente de estos métodos es su alto coste computacional, ya que es necesario que vayan perturbando la entrada de manera secuencial. Además, el modo en que se va a perturbar la entrada también puede cambiar mucho el resultado final, y no está claro cómo se puede hacer esto de manera óptima. Como resultado del experimento se puede descubrir la causa de alguna decisión puntual, pero es difícil interpretar toda la lógica interna del modelo.

Otras técnicas populares son las técnicas de maximización de la activación: aquellos métodos que crean artificialmente de manera iterativa una imagen de entrada para ilustrar el máximo estímulo de una neurona específica de cualquier capa modelo IA (Simonyan et al., 2013; Mahendran y Vedaldi, 2014; Yosinski et al., 2015). Un ejemplo se muestra en la figura 3.

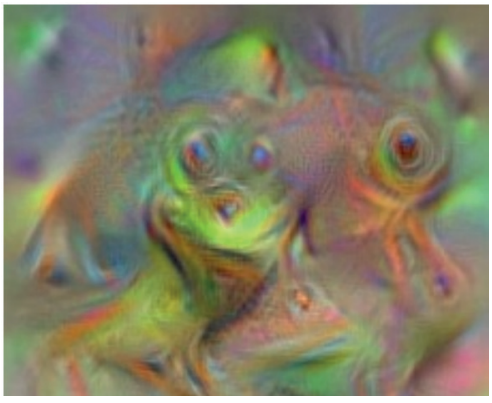

Figura 3. Imagen artificialmente generada que representa la activación máxima de una neurona entrenada para distinguir imágenes de ranas. Imagen extraída de (Mahendran y Vedaldi, 2014).

Existen también técnicas basadas en la retropropagación (*back-propagation*): normalmente calculan el gradiente (o alguna de sus variantes) de una salida en particular con respecto a la entrada para derivar la contribución de las características, dando lugar a mapas que indican qué partes de la imagen de entrada son más relevantes, los denominados mapas de saliencia (*saliency maps*) (Simonyan et al., 2013). Un ejemplo se muestra en la figura 4.

Figura 4. Mapa de saliencia donde se indica qué partes de la imagen son importantes para que la red pueda clasificar adecuadamente un objeto determinado (un perro en este ejemplo). Imagen extraída de (Simonyan et al., 2013).

Tanto los métodos de interpretación basados en la perturbación como los basados en back-propagation ignoran las capas intermedias de la red, que podrían contener información importante sobre la interpretabilidad de la misma. Hay otro tipo de métodos que investigan lo que ocurre en las capas ocultas de la red, para intentar determinar qué características de la entrada son más relevantes. Dentro de estos métodos podemos citar las técnicas GradCAM (Selvaraju et al., 2017) y ScoreCAM (Wang et al., 2019). Estas técnicas generan un mapa de calor sobre la imagen original, marcando las zonas más relevantes de la imagen para la decisión final del modelo. Estos mapas de calor (*heatmaps*) atribuyen una importancia a cada píxel de la imagen analizando la activación de la última capa convolucional del modelo (figura 5).

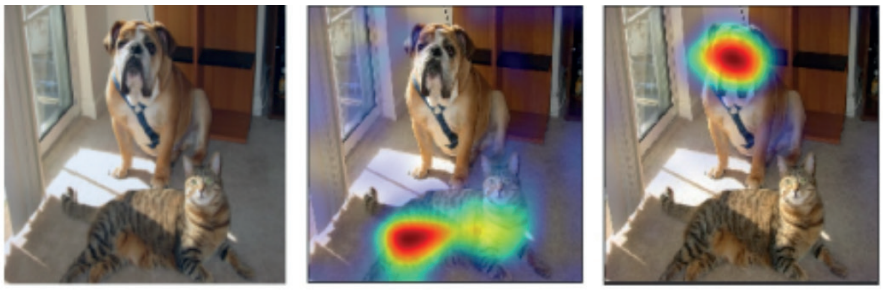

Figura 5. Mapas de calor (Grad-CAMs) que indican qué partes de la imagen son importantes para la red cuando detecta un gato o un perro. Imagen extraída de (Selvaraju et al., 2017).

Por último, recientemente han surgido técnicas que combinan el modelo IA clasificador (que toma la decisión) con otro modelo IA generador. El modelo generador tiene como objetivo generar entradas de manera artificial para poner a prueba al modelo clasificador. Estas técnicas pueden ayudar a descubrir si el modelo clasificador ha aprendido a distinguir alguna etiqueta por razones incorrectas (Shen et al., 2024; Sheng-Min et al., 2021).

Estas técnicas existentes son insuficientes y a veces no son coherentes. Por ejemplo, en el caso de modelos IA que trabajan con imágenes, muchas veces las técnicas de interpretabilidad parecen indicar que estos modelos IA fallan porque se fijan en sitios de la imagen irrelevantes, pero también hay disparidad de resultados si cambiamos de técnica de interpretación. Ante esto, surge una duda importante: ¿qué es incorrecto, el modelo de IA o el método que intenta interpretar lo que hace este modelo? Está claro que es necesaria más investigación enfocada a la interpretación interna de los modelos de IA, puesto que el problema no es trivial.

A todo esto, hay que añadir que los métodos actualmente utilizados presentan otro gran inconveniente: sólo son capaces de descubrir en qué características de más bajo nivel se fija el sistema, lo que resulta insuficiente para comprender bien las decisiones que toma, y, por lo tanto, este tipo de explicación no está orientada al usuario final (Du et al., 2019). Por este motivo, los investigadores están trabajando en el desarrollo de nuevos métodos de interpretabilidad en la IA que, además de incluir técnicas de visualización de redes neuronales, sean capaces de ofrecer una explicación de los resultados de la IA en lenguaje natural. Si bien estos enfoques son los más deseables, todavía hay mucho trabajo por hacer para lograr una interpretabilidad completa y confiable en la IA.

4.3. Un ejemplo claro: el problema de la interpretabilidad en las aplicaciones de IA diseñadas para el entorno médico

El entorno médico fue uno de los primeros campos donde se constató la necesidad real de dotar de explicabilidad a los sistemas IA. Los especialistas mé-

dicos actualmente no se niegan a usar herramientas IA, pero siempre bajo supervisión: no se fían del todo. En el ámbito médico, cualquier decisión conlleva un riesgo. Un médico analiza cuidadosamente los síntomas y los exámenes médicos del paciente antes de decidir si el paciente está enfermo o no, decisión que acompaña siempre de una explicación. Por tanto, para ser una herramienta viable y aceptada, la IA debe imitar el juicio y la capacidad de interpretación humanos.

Actualmente hay múltiples aplicaciones IA publicadas en la literatura científica para realizar diagnóstico analizando imágenes de pruebas médicas que no llegan a integrarse en la clínica diaria porque, como se ha indicado anteriormente, los especialistas médicos no se fían de ellas. Parecen ser sistemas de diagnóstico muy eficientes, pero actúan como cajas negras, y no dan explicaciones de los motivos que las han llevado a tomar la decisión final. Es necesario implementar estrategias de explicabilidad en estos sistemas por diversos motivos:

- Es esencial asegurarse de que el sistema se está fijando realmente en la información relevante de la enfermedad y no en otros detalles ajenos a la misma.
- Sería muy enriquecedor comprender o visualizar qué factores o características de alto nivel extrae el sistema de la información de entrada para tomar decisiones. Este conocimiento podría reforzar el del especialista médico, e incluso desvelar nuevos aspectos influyentes en el diagnóstico hasta ese momento desconocidos.
- La integración completa de este tipo de sistemas en la actividad asistencial sólo ocurrirá si el sistema se gana la confianza del usuario final (el médico), y eso implica hacer más transparente el comportamiento interno del sistema, ofreciendo una explicación razonada de las decisiones tomadas, tal y como lo hace un médico en su rutina diaria (una explicación "amigable", lo que en inglés se denomina "human-friendly").

La crisis sanitaria del COVID-19 despertó las alarmas en este sentido. Desde que se apreció la gravedad del problema, los investigadores de IA comenzaron a colaborar con los radiólogos de los hospitales para tratar de diseñar sistemas de IA que fueran capaces de detectar con precisión COVID-19 en radiografías de tórax. Muchos trabajos se publicaron al respecto en un corto periodo de tiempo (Han et al., 2023). Sin embargo, la solidez de estos sistemas se puso en duda posteriormente: algunos autores, utilizando técnicas de IA explicable, demostraron que muchos de estos sistemas se basaban en factores confusos en lugar de en la patología médica para tomar decisiones, creando una situación alarmante en la que los sistemas parecen precisos, pero fallan cuando se prueban en nuevos hospitales (DeGrave et al., 2021). En (Majeed et al, 2020) se utilizaron técnicas gráficas para destacar las zonas de las imágenes de rayos X que más influían en la decisión final de doce sistemas de IA diferentes entrenados para el diagnóstico de COVID-19. Debido a las incoherencias y a la gran disparidad encontrada, llegaron a la conclusión de que las decisiones de las CNN no deben tomarse en consideración, a pesar de su elevada precisión de clasificación, hasta que los

clínicos pudieran inspeccionar visualmente y aprobar la región o regiones de la imagen de entrada utilizadas por los sistemas IA que conducen a su predicción.

Los investigadores de IA que desarrollan aplicaciones para el campo sanitario son ahora más críticos con los resultados de los modelos de IA inexplicables, y orientan sus esfuerzos a métodos interpretables (Markus et al., 2021; Chaddad et al., 2023; Amann et al., 2020). En este campo se asume ya que la IA explicable debe considerarse un requisito previo para la implantación clínica de modelos sanitarios de aprendizaje automático.

4.4. Problemas de interpretabilidad en aplicaciones de IA diseñadas para la Administración Pública

En el ámbito de la Administración también podemos encontrar aplicaciones de IA fuerte que han originado problemas en su implantación por culpa de su escasa interpretabilidad. Un caso digno de mencionar, que tuvo gran trascendencia, fue el programa de IA "System Risk Indication" - SyRI (en español podríamos traducirlo como "Indicador de riesgo del sistema"), puesto en marcha en el año 2014 por el gobierno holandés (Algorithm Watch, 2024). Su objetivo era detectar el fraude en las prestaciones sociales. Sin embargo, a principios de 2020, el tribunal de distrito de La Haya ordenó la suspensión inmediata del programa.

A SyRI se le permitía cruzar datos sobre trabajo, multas, sanciones, impuestos, propiedades, vivienda, educación, jubilación, deudas, beneficios, asignaciones, subsidios, permisos y exenciones, etc. Estos datos los obtenía de una amplia gama de administraciones públicas. A partir de ellos, el sistema generaba "informes de riesgo", notificando cuándo había probabilidades de que se cometiesen irregularidades, y señalando los sujetos que consideraba "de riesgo" y que debían ser objeto de investigación.

A pesar de que el sistema contaba con múltiples garantías, como de anonimización, confidencialidad, etc., el tribunal, tras examinar a fondo su funcionamiento, dictó una sentencia, en la que, a pesar de reconocer los esfuerzos de la Administración Pública por ofrecer ciertas garantías en el uso de la aplicación, determinó que su empleo resultaba contrario al artículo 22 RGPD y al artículo 8 del Convenio Europeo de Derechos Humanos, ya que había carencias como la ausencia de auditorías externas independientes sobre el funcionamiento del algoritmo y la posible incidencia en derechos fundamentales de los ciudadanos, así como problemas derivados de la opacidad o "caja negra" del sistema, esto es, falta de transparencia y explicabilidad del algoritmo en que se basaba el sistema de IA implantado por la Administración (Berning, 2023).

Otro caso bastante mediático es el uso, desde el año 2000, de un programa de IA en EEUU denominado "Correctional Offender Management Profiling for Alternative Sanctions" – COMPAS (en español puede traducirse como "Administración de Perfiles de Criminales para Sanciones Alternativas"). Cuando una

persona es arrestada debe contestar un cuestionario. Analizando las respuestas de este cuestionario, el programa COMPAS indica si esa persona en el futuro podría cometer un crimen, y ese resultado lo tiene en cuenta el juez a la hora de juzgar el delito. Los investigadores y defensores de los derechos civiles llevan años dudando públicamente de su fiabilidad, puesto que se trata de un producto comercial cerrado, totalmente opaco, que no ofrece explicaciones de sus decisiones, y que, por tanto, no garantiza que sus resultados sean justos. Un estudio afirma haber demostrado que COMPAS no es más preciso o justo que las predicciones realizadas por personas con poca o ninguna experiencia en justicia penal, y que, además, está sesgado desde un punto de vista racial. Aunque los datos utilizados por COMPAS no incluyen la raza del individuo, otros aspectos de los datos pueden estar correlacionados con la raza, lo que puede dar lugar a disparidades raciales en las predicciones (Dressel y Farid, 2018).

5. LA CONFIANZA EN LA IA POR PARTE DE LOS USUARIOS

La confianza en la IA es un tema que preocupa a la Unión Europea (UE) (Comisión Europea, 2021). Como ya hemos comentado anteriormente, no poder entender bien cómo el modelo de IA funciona internamente (es decir, interpretarlo) genera inseguridad en los investigadores. Adicionalmente, que no se pueda garantizar siempre el buen uso de la IA es lo que preocupa a la mayoría de los ciudadanos.

La UE, consciente de este temor, ha diseñado un nuevo marco jurídico para garantizar que los sistemas de IA utilizados en la UE sean seguros, transparentes, éticos e imparciales, y estén bajo control humano. En este sentido, Europa es pionera. El Pleno del Parlamento Europeo acaba de ratificar la **primera ley de Inteligencia Artificial** del mundo, que aún deberá ser refrendada por el Consejo de la Unión Europea en las próximas semanas, antes de su entrada en vigor previsiblemente en el año 2026 (RTVE, 2024). Según esta norma, la UE etiqueta los sistemas IA en una de las siguientes categorías (figura 6):

- Riesgo inaceptable: aplicaciones de IA prohibidas, porque son una amenaza clara para los ciudadanos de la UE.
- Riesgo alto: sistemas de IA que se integren en actividades decisivas para los ciudadanos, como, por ejemplo, el transporte, la educación, la salud, la justicia, etc. En esta categoría se incluyen todos los servicios públicos, y, por tanto, toda la actividad de la Administración Pública en la que intervenga la IA fuerte.
- Riesgo reducido: sistemas de IA débil, como los robots conversacionales (chatbots), que están sujetos a unas obligaciones mínimas de transparencia en su interacción con los usuarios.
- Riesgo mínimo: aplicaciones basadas en IA débil que no representan riesgo alguno para los usuarios (videojuegos, filtros de correo no deseado, etc.).

La normativa de la UE se aplicará a los sistemas IA fuerte, por ser de riesgo alto.

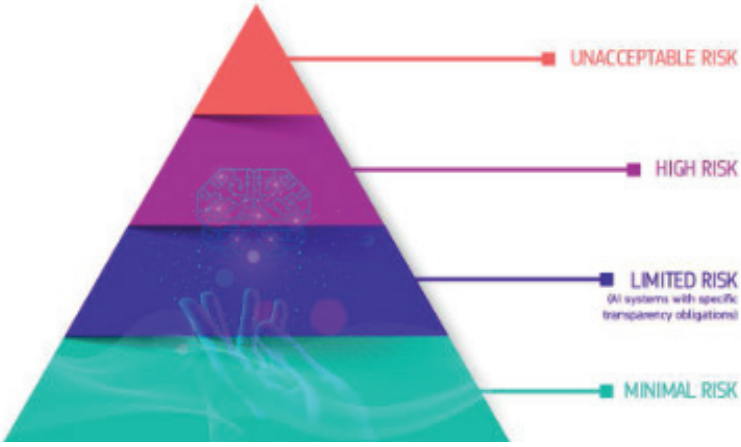

Figura 6. Clasificación de los sistemas de IA según la nueva normativa europea. Imagen extraída de (Comisión Europea, 2021).

La prioridad de esta ley es garantizar que *"los sistemas de IA utilizados en la UE sean seguros, transparentes, trazables, no discriminatorios y respetuosos con el medio ambiente"* (Parlamento Europeo, 2024). Resulta interesante que en esta legislación se contempla poner a disposición de las entidades desarrolladoras de IA espacios controlados de pruebas y ensayos en condiciones reales a nivel nacional, para que puedan desarrollar y entrenar la IA innovadora antes de su comercialización, como impulso a la innovación, y para facilitar que cumplan con los requisitos prioritarios mencionados anteriormente.

El problema de la interpretabilidad es uno de los primeros reconocidos por la propia UE, que manifiesta que: *"la opacidad de muchos algoritmos puede crear incertidumbre y obstaculizar la aplicación efectiva de la legislación vigente en materia de seguridad y derechos fundamentales. Para hacer frente a estos retos, es necesaria una acción legislativa que garantice el correcto funcionamiento del mercado interior de los sistemas de inteligencia artificial, con una ponderación adecuada de los beneficios y de los riesgos"* (Comisión Europea, 2024), por lo que la ley impone obligaciones específicas de transparencia que deben cumplir los sistemas IA.

Tal y como indica (Ortiz de Zárate, 2022), una IA fiable debe ser legal, ética y robusta. El primer requisito, que sea legal, ha sido un gran impedimento para su aplicación práctica en muchos entornos, puesto que hasta ahora no existía una normativa oficial aplicable. Esperemos que esta primera ley constituya realmente un marco de desarrollo de aplicaciones IA dignas de confianza.

6. APLICACIONES DE IA ACTUALMENTE ACTIVAS EN LA ADMINISTRACIÓN PÚBLICA ESPAÑOLA

Hoy en día existen varias aplicaciones de IA operando en nuestra Administración Pública. Se trata, mayoritariamente, como ya hemos mencionado, de aplicaciones de IA débil, que actúan de manera controlada en entornos restringidos, y que son capaces de automatizar ciertas tareas rutinarias propias de la administración, pero podemos encontrar también ejemplos de aplicaciones IA con más autonomía, de IA fuerte.

Una de las aplicaciones más extendidas son los "chatbots" o asistentes virtuales: sistemas IA que se utilizan para asistir telemáticamente a los ciudadanos de forma automatizada, de forma que no es necesaria la presencia de personas, permitiendo que una consulta sencilla pueda ser respondida a cualquier hora del día o la noche, sea festivo o no. Estos programas han mejorado mucho gracias al avance de las técnicas de Procesamiento de Lenguaje Natural (NLP). A pesar de ser herramientas útiles, su empleo debe tener ciertos límites, porque se corre el riesgo de descuidar la atención al ciudadano, como han puesto de manifiesto algunos expertos (Berning, 2023).

Otra aplicación IA muy interesante y útil son los formularios inteligentes, que automatizan la cumplimentación de formularios, como, por ejemplo, los usados para hacer autodeclaraciones de la renta (programa Renta Web). Este programa es de gran ayuda para el ciudadano, que únicamente se debe preocupar de comprobar que los datos proporcionados son correctos y añadir aquellos no incluidos (Berning, 2023). La Agencia Estatal de Administración Tributaria tiene más servicios automatizados, actividades de mero trámite. Aunque es destacable la herramienta de IA que puso en funcionamiento en 2020 para analizar la gran cantidad de datos (big data) disponible en sus servidores para vigilar, de forma automática, las operaciones internacionales de las multinacionales (Faes, 2020).

Hay administraciones públicas que usan la IA para tratar de personalizar los servicios públicos. Para ello, elaboran perfiles sobre el comportamiento de los usuarios y analizan sus datos personales. Un ejemplo de esto es el Recomendador de Ayudas Sociales "MyGov Social", de la Administració Oberta de Catalunya (Administració Oberta de Catalunya, 2023).

Las administraciones públicas también usan la IA para detectar fraudes y casos de corrupción. Como ejemplo, podemos citar el Sistema de alerta rápida SALER, un sistema de alertas contra la corrupción de la Administración de la Comunidad Valenciana (Anti-Fraud Knowledge Centre, 2021).

La Policía Nacional usa desde hace unos años el programa VeriPol, una aplicación de IA que a partir de técnicas de Procesamiento del Lenguaje Natural (NLP) y Aprendizaje Automático (ML), ayuda a detectar denuncias falsas. El sistema estudia las expresiones y palabras utilizadas al redactar las denuncias para inferir cuándo la persona denunciante miente. Se trata de la primera herramienta de este tipo en el mundo y, con una precisión de más del 90%, estima la

probabilidad de que una denuncia por robo con violencia e intimidación o tirón sea falsa (Policía Nacional, 2018).

Por último, mencionar el proyecto europeo "JuLIA: Justice, fundamental rights and artificial intelligence", coordinado por la Universidad Pompeu Fabra de Barcelona, en el que se analiza cómo integrar la IA en las decisiones automatizadas en la labor jurisdiccional. Los avances del proyecto se pueden consultar en (Universitat Pompeu Fabra, 2021).

CONCLUSIONES

La integración de la inteligencia artificial (IA) en la Administración es inevitable, como parte del proceso de transición digital en el que estamos inmersos. Se avecina una verdadera revolución en los servicios públicos. Las ventajas son múltiples, como analizaremos a continuación, pero los riesgos, por desgracia, también (Huergo, 2023).

Comencemos analizando las ventajas, porque son importantes, y es lo que justifica que los gobiernos estén actualmente invirtiendo tanto en el desarrollo de IA para la Administración. La IA debe servir para:

- Reducir el coste de los servicios públicos.
- Facilitar la toma de decisiones.
- Agilizar los tiempos de los procesos involucrados.
- Implementar políticas más éticas, confiando en que las máquinas sean más imparciales.
- Conseguir mayor precisión ante situaciones complejas.
- Facilitar el trabajo de los recursos humanos de la Administración, liberarlos de tareas rutinarias para que puedan invertir su tiempo en actividades más importantes.

Pero esta integración de la IA en la Administración conlleva bastantes dificultades y exigencias (Boix, 2022). En el escenario actual, se han presentado diversas iniciativas basadas en nuevas tecnologías que parecen prometedoras para su aplicación en distintos ámbitos de la Administración Pública. Sin embargo, la experiencia ha demostrado que no todas estas soluciones son adecuadas y que no todas las soluciones adecuadas son aplicables en cualquier contexto (Etscheid, 2019).

La primera dificultad que analizaremos será el **problema de la interpretabilidad**, por ser el objetivo del presente artículo. Uno de los pilares esenciales del Estado de Derecho es la transparencia y publicidad de la actuación administrativa. Los actos de las administraciones públicas deben ser motivados y públicos, totalmente transparentes, de tal forma que cualquier ciudadano pueda saber las razones de la decisión (Ley 39/2015, de 1 de octubre, del Procedimiento Admi-

nistrativo Común de las Administraciones Públicas). Por este motivo, el requisito de explicabilidad para una aplicación de IA es imperativo.

Como se ha expuesto en este artículo, el problema de la interpretabilidad de la IA es un tema actualmente bajo investigación. Muchas IA son desarrolladas por empresas privadas que no permiten que personas externas vean cómo funcionan (Dressel y Farid, 2018). Por otra parte, los modelos subyacentes en estas aplicaciones están basados en redes neuronales artificiales tan complejas que ni los propios diseñadores cuentan con las herramientas necesarias para desentrañar su funcionamiento interno, lo que ha disparado las alarmas y generado preocupación. La IA avanza a un ritmo vertiginoso, y no somos capaces de entender o controlar este crecimiento de sistemas de "caja negra". Esta situación ha empujado a Europa a elaborar la primera ley de IA mundial, que intenta establecer un marco legal para el desarrollo de aplicaciones IA en todos los ámbitos, incluido el sector público.

Este desconocimiento del comportamiento de los modelos de IA fuerte genera desconfianza en los usuarios. Puede que a una persona no le importe que una IA le haga alguna recomendación banal o le atienda en un chatbot, pero si está en juego que le concedan una subvención, le apliquen una sanción, o incluso, una pena de prisión, es normal que surja la desconfianza, si la aplicación IA que resuelve su caso no es capaz de explicarle los motivos de la decisión, que es algo básico en todo Estado de Derecho para poder garantizar su derecho de defensa. En mi opinión, este es el mayor obstáculo para la integración de la IA fuerte en la Administración Pública.

Además, la transparencia en los sistemas IA es fundamental para prevenir la discriminación y garantizar la equidad en la toma de decisiones. Los algoritmos pueden contener sesgos si no se entrenan adecuadamente (recordemos la importancia de los datos, mencionada en este artículo). Esta característica, sumada a la opacidad del algoritmo, supone que en muchas ocasiones las discriminaciones sean difícilmente detectables (Dressel y Farid, 2018; Mendilibar, 2023).

Haciendo hincapié en **la importancia de los datos**: la eficiencia y la equidad de los sistemas de IA depende de la cantidad y la calidad de los datos con los que se entrenan. Esto implica que la adopción de sistemas de IA en una administración pública depende de la capacidad de un gobierno para recopilar, almacenar, procesar y compartir datos entre las diferentes organizaciones que lo componen (Filgueiras, 2021), y esto, en muchos casos representa una dificultad añadida. Hay que garantizar también la protección de los datos, así como el derecho de la supresión de la información de aquellas personas que no quieran ser objeto de decisiones algorítmicas (Mendilibar, 2023; Algorithm Watch, 2024).

Por otro lado, pueden surgir dificultades en el proceso de integración de la IA en la Administración relacionadas con **los empleados públicos**. El objetivo de la IA en no es sustituir al personal de la administración. Al contrario, la IA debe tener como principal y única finalidad empoderar a los diferentes roles y

trabajadores que intervienen en la gestión, dotándolos de nuevas herramientas, nuevos cuadros de mando, nuevas habilidades de gestión y comunicación que permitan que su trabajo diario sea mucho más eficiente y rápido, y que aporte valor real a la ciudadanía. Pero es evidente que sus competencias van a cambiar, y esto implica que su formación también, tendrán que aprender a manejar nuevas tecnologías, y se especializarán en aquellas competencias que la IA no es capaz de realizar. Este proceso de adaptación puede no ser bien recibido por todos los empleados, que pueden oponerse a la transición digital. En este sentido, la crisis sanitaria originada por el COVID-19 aceleró esta transición en el sector público, siendo un punto de inflexión en la adquisición de competencias digitales del personal de dicho sector (Mendilibar, 2023).

Desde el punto de vista técnico, es fundamental considerar la estructura organizativa, los procesos existentes, la capacitación del personal y las necesidades de los usuarios finales. Pero comprender cuáles son los actores que participan en las decisiones de la Administración Pública y cómo lo hacen no es una tarea sencilla. La administración pública no tiene un protocolo establecido para diseñar sistemas de IA, y existen dificultades para comunicar las demandas entre los tomadores de decisiones y los desarrolladores de IA. Por lo tanto, el diseño de las soluciones de los sistemas IA podría no coincidir con el diseño de las soluciones de los problemas de la administración pública, por un **problema de comunicación** entre profesionales de distintas disciplinas (Filgueiras, 2021).

Diversos autores sugieren algunas actuaciones para intentar minimizar los efectos negativos de la integración de la IA en la administración (Mendilibar, 2023; Berning, 2023; Ponce, 2019):

- Realización de auditorías algorítmicas y evaluaciones normativas.
- Realización de auditorías sobre los datos, que permitan conocer y controlar los datos que se van a usar en el diseño de la IA, para garantizar así su calidad.
- Permitir la participación de terceros en el procedimiento de aprobación de la aplicación.
- Hacer una *reserva de humanidad* en casos concretos y especialmente sensibles, es decir, que se reserve la toma de determinadas decisiones a los humanos en esos casos.
- Realización de controles de los actos basados en IA.
- Aumentar el número de técnicos especializados en materia de IA.

A todas estas actuaciones me gustaría añadir, desde mi punto de vista técnico, que es necesario dedicar más tiempo y más inversión a la investigación sobre la interpretabilidad de los modelos IA fuerte. Muchos de los problemas mencionados anteriormente no serían tales si fuéramos capaces de explicar el comportamiento interno de estos sistemas.

En resumen, en el diseño de nuevas soluciones de IA para la administración se debe considerar no solo la tecnología en sí misma, sino también aspectos como la viabilidad técnica, la capacidad de integración con los sistemas existentes, la formación y adaptación del personal, la sostenibilidad a largo plazo de las soluciones propuestas, y, como se ha explicado en este artículo, el grado de interpretabilidad o explicabilidad. Solo mediante una combinación equilibrada y bien planificada de tecnologías y de estos aspectos relevantes se podrán obtener soluciones viables y eficaces que realmente puedan añadir valor al servicio final ofrecido por la Administración Pública, y que además cumplan con **seis aspectos clave**:

- **Imparcialidad**: sistemas que traten a todos los seres humanos por igual. Eso significa eliminar los sesgos y la discriminación.
- **Responsabilidad**: se debe determinar a quién se le atribuye la responsabilidad del uso de estos sistemas.
- **Transparencia**: sistemas donde es posible entender cómo funcionan los algoritmos que contienen y que ofrezcan una explicación comprensible de las razones por las que toman determinadas decisiones.
- **Fiabilidad**: sistemas fiables y robustos, con resultados reproducibles y coherentes.
- **Seguridad:** sistemas que garanticen que los datos sensibles se almacenen y utilicen de forma segura y confidencial.
- **Ética**: sistemas entrenados para respetar la ética y los valores importantes de la sociedad, y cumplir la normativa aplicable.

BIBLIOGRAFÍA

(Moor, 2006) Moor J. "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years" (2006). AI Magazine, 27:87-91.

(Rouse, 2024a) Rouse M. "Inteligencia artificial débil" (2024). Disponible online: https://www.techopedia.com/es/definicion/inteligencia-artificial-debil. Último acceso: 17/03/24.

(Rouse, 2024b) Rouse M. "Inteligencia artificial fuerte" (2024). Disponible online: https://www.techopedia.com/es/definicion/inteligencia-artificial-fuerte. Último acceso: 17/03/24.

(Kim et al., 2022) Kim I., Kang K., Song Y., Kim T.J. "Application of Artificial Intelligence in Pathology: Trends and Challenges" (2022). Diagnostics (Basel), 15;12(11):2794. https://doi.org/10.3390/diagnostics12112794

(Kumar et al., 2023) Kumar Y., Koul A., Singla R., Ijaz M.F. "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda" (2023). J Ambient Intell Humaniz Comput, 14(7):8459-8486. https://doi.org/10.1007/s12652-021-03612-z

(Yala et al., 2019) Yala A., Lehman C., Schuster T., Portnoi T., Barzilay R. "A deep learning mammography-based model for improved breast cancer risk prediction" (2019). Radiology, 292(1), 60-66. https://doi.org/10.1148/radiol.2019182716

(Zhou et al., 2023) Zhou X., Chen Y., Ip F.C.F. et al. "Deep learning-based polygenic risk analysis for Alzheimer's disease prediction" (2023). Commun Med 3, 49. https://doi.org/10.1038/s43856-023-00269-x

(Gallent et al., 2023) Gallent C., Zapata A., Ortego J.L. "El impacto de la inteligencia artificial generativa en educación superior: una mirada desde la ética y la integridad académica" (2023). Relieve. 29, 2. https://doi.org/10.30827/relieve.v29i2.29134

(Franganillo, 2023) Franganillo J. "La inteligencia artificial generativa y su impacto en la creación de contenidos mediáticos" (2023). Methaodos: revista de ciencias sociales, 11 (2).

(Linardatos et al., 2020) Linardatos P., Papastefanopoulos V., Kotsiantis S. "Explainable AI: A Review of Machine Learning Interpretability Methods" (2020). Entropy (Basel), 25;23(1):18. doi: 10.3390/e23010018.

(Mariano, 2002) Mariano J.A. "Automatización de la gestión de expedientes administrativos" (2002). VII Jornadas sobre Tecnologías de la Información para la modernización de las Administraciones Públicas. Ponencia.

(Rodríguez y González, 2002) Rodríguez J.V., González J. "Integración de las tecnologías de flujo de trabajo y gestión documental para la optimización de los procesos de negocio" (2002). Ciencias de la Información, 33(3), pp. 17.

(Comisión Europea, 2019) Comisión Europea. "Prioridades de la Comisión Europea 2019-2024". Disponible online: https://spain.representation.ec.europa.eu/estrategias-y-prioridades/prioridades-de-la-comision-europea-2019-2024_es. Último acceso: 17/03/24.

(Gobierno de España, 2024a) Gobiernto de España. "España Digital 2026". Disponible online: https://portal.mineco.gob.es/en-us/ministerio/estrategias/Pages/00_Espana_Digital.aspx. Último acceso: 17/03/24.

(Gobierno de España, 2024b) Gobierno de España. "Catálogo de servicios de Administración digital". Disponible online: https://administracionelectronica.gob.es/pae_Home/pae_Estrategias/Racionaliza_y_Comparte/catalogo-servicios-admon-digital.html. Último acceso: 17/03/24.

(Veale y Brass, 2019) Veale M., Brass I. "Administration by Algorithm? Public Management Meets Public Sector Machine Learning. Algorithmic Regulation" (2019). Oxford University Press.

(Restrepo-Amariles, 2020) Restrepo-Amariles D. "Algorithmic Decision Systems: Automation and Machine Learning in the Public Administration" (2020) The Cambridge Handbook of the Law of Algorithms.

(Anastasopoulos y Whitford, 2019) Anastasopoulos L.J., Whitford A.B. "Machine Learning for Public Administration Research, With Application to Organizational Reputation" (2019). Journal of Public Administration Research and Theory, 29(3), 491–510. https://doi.org/10.1093/jopart/muy060

(Henman, 2020) Henman P. "Improving public services using artificial intelligence: possibilities, pitfalls, governance" (2020). Asia Pacific Journal of Public Administration, 42:4, 209-22.

(Cerrillo i Martínez, 2019) Cerrillo i Martínez A. "El impacto de la inteligencia artificial en el derecho administrativo ¿nuevos conceptos para nuevas realidades técnicas?" (2019). Revista General de Derecho Administrativo, núm. 50.

(Etscheid, 2019) Etscheid J. "Artificial Intelligence in Public Administration" (2019). In: Electronic Government. EGOV 2019. Lecture Notes in Computer Science, vol 11685. Springer, Cham.

(Sobrino-García, 2021) Sobrino-García I. "Artificial Intelligence Risks and Challenges in the Spanish Public Administration: An Exploratory Analysis through Expert Judgements" (2021). Administrative Sciences, 1; 11(3):102.

(Cabanillas et al., 2012) Cabanillas C., Resinas M., Ruiz-Cortés A. "Automated Resource Assignment in BPMN Models Using RACI Matrices (2012). In: On the Move to Meaningful Internet Systems OTM 2012. Lecture Notes in Computer Science, vol 7565. Springer, Berlin, Heidelberg.

(Mullakara y Asokan, 2020) Mullakara N., Asokan A.K. "Robotic Process Automation Projects: Build real-world RPA solutions using UiPath and Automation Anywhere" (2020) Ed. Packt Publishing.

(Houy et al., 2019) Houy C., Hamberg M., Fettke P. "Robotic Process Automation in Public Administrations" (2019). Conference: Digitalisierung von Staat und Verwaltung. Münster, Germany.

(Uskenbayeva et al., 2019) Uskenbayeva R., Kalpeyeva Z., Satybaldiyeva R., Moldagulova A., Kassymova A. "Applying of RPA in Administrative Processes of Public Administration" (2019). IEEE 21st Conference on Business Informatics (CBI), 9-12. 10.1109/CBI.2019.10089.

(Johansson et al., 2022) Johansson J., Thomsen M., Åkesson M.A. "Public value creation and robotic process automation: normative, descriptive and prescriptive issues in municipal administration" (2022). Transforming Government: People, Process and Policy.

(Kang et al., 2020) Kang Y., Cai Z., Tan C.W., Huang Q., Liu H. "Natural language processing (NLP) in management research: A literature review" (2020). Journal of Management Analytics, 7:2, 139-172.

(Kowalski et al., 2017) Kowalski R., Esteve M., Mikhaylov S. "Application of Natural Language Processing to determine user satisfaction in Public Services" (2017). arXiv:1711.08083. https://doi.org/10.48550/arXiv.1711.08083.

(Comisión Europea, 2022) Comisión Europea. "Natural Language Processing for Public Services" (2022). Disponible online: https://joinup.ec.europa.eu/sites/default/files/inline-files/D02.01_Natural%20Language%20Processing%20for%20Public%20Services_4.pdf. Último acceso: 17/03/2024.

(Chen, 2016) Chen C.L.P. "Big Data challenges, techniques, technologies, and applications and how deep learning can be used" (2016). IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Nanchang, China, pp. 3-3.

(Sarker, 2021) Sarker, I.H. "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions" (2021). *SN COMPUT. SCI.* 2, 420. https://doi.org/10.1007/s42979-021-00815-1.

(Torres, 2020) Torres J. "Python Deep Learning. Introducción práctica con Keras y Tensorflow 2" (2020). Ed. Marcombo.

(Felderer y Ramler, 2021) Felderer M., Ramler R. "Quality Assurance for AI-Based Systems: Overview and Challenges" (2021). In: Winkler, D., Biffl, S., Mendez, D., Wimmer, M., Bergsmann, J. (eds) Software Quality: Future Perspectives on Software Engineering Quality. SWQD 2021. Lecture Notes in Business Information Processing, vol 404. Springer, Cham. https://doi.org/10.1007/978-3-030-65854-0_3.

(Martens, 2018) Martens, B. "The Importance of Data Access Regimes for Artificial Intelligence and Machine Learning" (2018). JRC Digital Economy Working Paper 2018-09, http://dx.doi.org/10.2139/ssrn.3357652.

(Gobierno de España, 2021) Gobierno de España. "Guía al Análisis Exploratorio de Datos. Ministerio de Asuntos Económicos y Transformación Digital". Disponible online:

https://datos.gob.es/es/documentacion/guia-practica-de-introduccion-al-analisis-exploratorio-de-datos. Último acceso: 17/03/24.

(Doshi-Velez y Kim, 2017) Doshi-Velez, F., y Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

(Saeed y Omlin, 2023) Saeed W., Omlin C. "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities" (2023) Knowledge-Based Systems, 263, 110273. https://doi.org/10.1016/j.knosys.2023.110273.

(Biran y Cotton, 2017) Biran O., Cotton C. "Explanation and justification in machine learning: A survey" (2017). In: IJCAI-17 Workshop on Explainable AI, Vol. 8, XAI, (1), 8–13.

(Gladstone AI, 2024) Gladstone AI. "An Action Plan to increase the safety and security of advanced AI" (2024). Disponible online: https://www.gladstone.ai/action-plan. Último acceso: 17/03/24.

(Woollacott, 2023) Woollacott E." Elon Musk y otros expertos en tecnología piden parar el entrenamiento de la Inteligencia Artificial". Forbes. Disponible online: https://forbes.es/tecnologia/256871/elon-musk-y-otros-expertos-en-tecnologia-piden-parar-el-entrenamiento-de-la-inteligencia-artificial/. Último acceso: 17/03/24.

(Han et al., 2023) Han X., Hu Z., Wang S., Zhang Y. "A Survey on Deep Learning in COVID-19 Diagnosis" (2023). J. Imaging, 9, 1. https://doi.org/10.3390/jimaging9010001

(DeGrave et al., 2021) DeGrave A.J., Janizek J.D., Lee, SI."AI for radiographic COVID-19 detection selects shortcuts over signal" (2021). Nat Mach Intell 3, 610–619. https://doi.org/10.1038/s42256-021-00338-7

(Majeed et al, 2020) Majeed T., Rashid R., Ali D. *et al.* "Issues associated with deploying CNN transfer learning to detect COVID-19 from chest X-rays" (2020). Phys Eng Sci Med 43, 1289–1303. https://doi.org/10.1007/s13246-020-00934-8

(Markus et al., 2021) Markus A.F., Kors J.A., Rijnbeek P.R. "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies" (2021). Journal of Biomedical Informatics, 113.

(Chaddad et al., 2023) Chaddad A., Peng J., Xu J., Bouridane A. "Survey of Explainable AI Techniques in Healthcare" (2023). Sensors (Basel), 5;23(2):634. doi: 10.3390/s23020634.

(Amann et al., 2020) Amann J., Blasimme A., Vayena E. *et al.* "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective" (2020). BMC Med Inform Decis Mak 20, 310. https://doi.org/10.1186/s12911-020-01332-6.

(Algorithm Watch, 2024) Algorithm Watch. "How Dutch activists got an invasive fraud detection algorithm banned". Disponible online: https://algorithmwatch.org/en/syri-netherlands-algorithm/. Último acceso: 17/03/24.

(Berning, 2023) Berning A.D. "El uso de sistemas basados en inteligencia artificial por las Administraciones públicas: estado actual de la cuestión y algunas propuestas ad futurum para un uso responsable" (20239. Revista de Estudios de la Administración Local y Autonómica (INAP), número 20.

(Dressel y Farid, 2018) Dressel J., Farid H. "The accuracy, fairness, and limits of predicting recidivism" (2018). Sci. Adv.4,eaao5580. DOI:10.1126/sciadv.aao5580

(Du et al., 2019) Du M., Liu N., Hu X. "Techniques for Interpretable Machine Learning" (2019). Communications of the ACM (CACM 2019). DOI:10.1145/3359786.

(Yosinski et al., 2015) Yosinski J., Clune J., Nguyen A., Fuchs T., Lipson H. "Understanding neural networks through Deep visualization (2015). In Deep Learning Workshop, ICML conference. arXiv:1506.06579. https://doi.org/10.48550/arXiv.1506.06579.

(Nguyen et al., 2016) Nguyen A., Yosinski J., Clune J. "Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks" (2016) In Visualization for Deep Learning Workshop, ICML conference, 2016.

(Ancona et al., 2018) Ancona M., Ceolini E., Öztireli C., Gross M. "Towards better understanding of gradient-based attribution methods for Deep Neural Networks" (2018). In International Conference on Learning Representations (ICLR 2018).

(Selvaraju et al., 2017) Selvaraju R.R., Das A., Vedantam R., Cogswell M., Parikh D., Batra D. "Grad-cam: visual explanations from deep networks via gradient-based localization" (2017). International Conference on Computer Vision (ICCV 2017). https://arxiv.org/abs/1610.02391

(Wang et al., 2019) Wang H., Du M., Yang F., Zhang Z. "Score-CAM: Improved Visual Explanations Via Score-Weighted Class Activation Mapping" (2019). eprint arXiv:1910.01279. Accepted to CVPR 2020: Workshop on Fair, Data Efficient and Trusted Computer Vision. https://arxiv.org/abs/1910.01279

(Zeiler y Fergus, 2013) Zeiler M.D., Fergus R. "Visualizing and Understanding Convolutional Networks" (2013). Computer Vision and Pattern Recognition. arXiv:1311.2901. https://doi.org/10.48550/arXiv.1311.2901.

(Bach et al., 2015) Bach S., Binder A., Montavon G., Klauschen F., Müller K.R., Samek W. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation" (2015). PLoS ONE 10(7): e0130140. https://doi.org/10.1371/journal.pone.0130140.

(Samek et al., 2015) Samek W., Binder A., Montavon G., Bach S., Müller K.R. ·Evaluating the visualization of what a Deep Neural Network has learned· (2015). Computer Vision and Pattern Recognition. arXiv:1509.06321. https://doi.org/10.48550/arXiv.1509.06321.

(Simonyan et al., 2013) Simonyan K., Vedaldi A., Zisserman A. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps" (2013).

Computer Vision and Pattern Recognition. arXiv:1312.6034. https://doi.org/10.48550/arXiv.1312.6034

(Mahendran y Vedaldi, 2014) Mahendran A., Vedaldi A. "Understanding Deep Image Representations by Inverting Them" (2014). arXiv:1412.0035. https://doi.org/10.48550/arXiv.1412.0035.

(Shen et al., 2024) Shen X., Song Z., Zhang Z. "AFBT GAN: enhanced explainability and diagnostic performance for cognitive decline by counterfactual generative adversarial network" (2024). arXiv preprint arXiv:2403.01758.

(Sheng-Min et al., 2021) Sheng-Min S., Tien P.J., Karnin Z. "GANMEX: One-vs-one attributions using GAN-based model explainability" (2021). International Conference on Machine Learning. PMLR.

(Comisión Europea, 2021) Comisión Europea. "Excelencia y confianza en la inteligencia artificial" (2021). Disponible online; https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/excellence-and-trust-artificial-intelligence_es. Último acceso: 17/03/24.

(RTVE, 2024) RTVE. "El Parlamento Europeo ratifica la primera ley de inteligencia artificial del mundo". Disponible online: https://www.rtve.es/noticias/20240313/parlamento-europeo-aprueba-primera-ley-inteligencia-artificial-del-mundo/16013050.shtml#:~:text=El%20Pleno%20del%20Parlamento%20Europeo,previsiblemente%20en%20el%20a%C3%B1o%202026. Último acceso: 17/03/24.

(Parlamento Europeo, 2024) Parlamento Europeo. "Ley de IA de la UE: primera normativa sobre inteligencia artificial" (2024) Disponible online: https://www.europarl.europa.eu/topics/es/article/20230601STO93804/ley-de-ia-de-la-ue-primera-normativa-sobre-inteligencia-artificial. Último acceso: 17/03/24.

(Comisión Europea, 2024) Comisión Europea. "Inteligencia artificial: preguntas y respuestas" (2024). Disponible online: https://ec.europa.eu/commission/presscorner/detail/es/QANDA_21_1683. Último acceso: 17/03/24.

(Ortiz de Zárate, 2022) Ortiz de Zárate L. "Explicabilidad (de la inteligencia artificial)" (2022). Eunomía. Revista en Cultura de la Legalidad, 22, 328-344. DOI: https://doi.org/10.20318/eunomia.2022.6819

(Faes, 2020) Faes I. "El 'big data' llega a Hacienda: un súperordenador vigilará a las multinacionales" (2020). Eleconomista.es. Disponible online: https://www.eleconomista.es/legislacion/noticias/10325315/01/20/El-big-data-llega-a-Hacienda-Un-superordenador-vigilara-a-las-multinacionales.html. Último acceso: 17/03/24.

(Administració Oberta de Catalunya, 2023) Administració Oberta de Catalunya. "Recomendador de ayudas sociales - MyGov Social" (2023), Disponible online: https://www.aoc.cat/es/projecte-innovacio/recomanador-dajuts-socials-mygov-social/. Último acceso: 17/03/24.

(Anti-Fraud Knowledge Centre, 2021) Anti-Fraud Knowledge Centre (UE). "Sistema de alerta rápida SALER" (2021). Disponible online: https://antifraud-knowledge-centre.ec.europa.eu/library-good-practices-and-case-studies/good-practices/saler-rapid-alert-system_es. Último acceso: 17/03/24.

(Policía Nacional, 2018) Policía Nacional. "La Policía Nacional pone en funcionamiento la aplicación informática VeriPol para detectar denuncias falsas" (2018). Disponible

online: https://www.policia.es/_es/comunicacion_prensa_detalle.php?ID=4433&idiomaActual=es. Último acceso: 17/03/24.

(Universitat Pompeu Fabra, 2021) Universitat Pompeu Fabra. "JULIA: Justice, Fundamental Rights and Artificial Intelligence" (2021). Disponible online: https://www.julia-project.eu/. Último acceso: 17/03/24.

(Huergo, 2023) Huergo A. "Inteligencia artificial: una aproximación jurídica no catastrofista" (2023). Revista Española de Control Externo, vol. XXV, n.º 74-75, pp. 110-129.

(Boix, 2022) Boix A. "Transparencia en la utilización de inteligencia artificial por parte de la Administración" (2022). El Cronista del Estado Social y Democrático de Derecho, núm. 100.

(Mendilibar, 2023) Mendilibar P. "Redefinición de las competencias de los empleados y empleadas públicas ante el uso de la Inteligencia Artificial por la Administración Pública. Revista Documentación Administrativa" (2023). INAP, número 10.

(Filgueiras, 2021) Filgueiras F. "Inteligencia Artificial en la administración pública: ambigüedad y elección de sistemas de IA y desafíos de gobernanza digital" (2021). Revista del CLAD Reforma y Democracia, No. 79, 5-38.

(Ponce, 2019) Ponce J. "Inteligencia artificial, Derecho administrativo y reserva de humanidad: algoritmos y procedimiento administrativo debido tecnológico". Revista General de Derecho Administrativo (Iustel), n. 50.