Inteligencia Artificial en las Administraciones Públicas: definiciones, evaluación de viabilidad de proyectos y áreas de aplicación.

La Inteligencia Artificial (IA) posee un indudable potencial y ya puede emplearse para numerosas aplicaciones. Sin embargo, para emprender con éxito un proyecto de IA, es esencial la comprensión de sus definiciones y limitaciones subyacentes; así como de los riesgos técnicos, organizativos y jurídicos que implica su uso. Trataremos de identificarlos y de proponer soluciones, teniendo en cuenta el grado de madurez actual de las tecnologías de IA y el marco jurídico aplicable a las Administraciones Públicas de España.

ALEXANDER ZLOTNIK
Jefe de Servicio. SGTI. Ministerio de Sanidad, Consumo y Bienestar Social.

Cuerpo Superior de Sistemas y Tecnologías de la Información de la Administración del Estado.

Ingeniero Superior de Telecomunicación (ETSIT-UPM) y Doctor en Inteligencia Artificial (UPM).

DEFINICIÓN INTUITIVA DE LA INTELIGENCIA ARTIFICIAL ACTUAL

La literatura académica recopila numerosas y variadas definiciones de la Inteligencia Artificial, desde los enfoques de las ciencias de computación, de las ciencias exactas y de las ciencias sociales [1, 2]. Sin embargo, puede ser más intuitiva una definición de *via negativa:* **la Inteligencia Artificial actual no es inteligencia**, en el sentido que los seres humanos otorgamos a esa palabra. Las técnicas de IA que han logrado los resultados más prometedores parten de un conjunto de variables de entrada y salida; la relación entre ellas se establece mediante un proceso de «entrenamiento» o «aprendizaje» realizado por **algoritmos**, generalmente guiados por grandes cantidades de datos. Dichos algoritmos de IA son capaces de construir soluciones (que llamaremos **modelos** de IA) a algunos problemas, pero sin una comprensión de la información subyacente. Es decir, *a priori* –y a no ser que lo indiquen sus creadores–, un algoritmo de IA asigna la misma relevancia a variables de entrada tales como el nivel de renta de un individuo, su código postal, su origen étnico o el color del cielo de un determinado día. Si dichas variables permiten cumplir el objetivo del algoritmo (p.ej. predecir una variable de salida), serán consideradas importantes por el algoritmo, aunque la información contenida en ellas sea espuria. Esto tiene varias implicaciones: (i) los modelos de IA pueden dar lugar a re-

sultados sesgados[1]; (ii) el rendimiento de los modelos puede decaer sensiblemente ante un cambio de condiciones, que modifique la interpretación de las variables[2]; (iii) el uso de numerosas variables[3], combinado o no con algoritmos de IA complejos, puede dar lugar modelos de IA difíciles de comprender para los humanos[4].

Por estos y otros motivos –que veremos a continuación–, **antes de plantearse una solución basada en IA, se debe realizar un estudio de alternativas, en el que se comparen los riesgos asumidos, así como los recursos y el tiempo necesarios para la construcción de una solución clásica, basada en programación tradicional (código fuente completamente transparente y predecible), con los de una solución basada en IA[5].**

DEFINICIONES DE CONCEPTOS RELACIONADOS CON LA IA

Aunque la definición intuitiva de IA del apartado anterior puede resultar aclaratoria, es importante mencionar además otras definiciones de uso común, tanto en el ámbito académico como en la industria.

En las ciencias de la computación *(computer science)*, se diferencia de manera clara la Inteligencia Artificial General (conocida como *general AI* o *strong AI*) –con capacidades similares a las de un ser humano–, de la Inteligencia Artificial Especializada (conocida como *narrow AI* o *weak AI*) –con capacidad para resolver únicamente problemas concretos–[2]. **Ningún sistema de IA actual puede considerarse una Inteligencia Artificial General (IAG) y, seguramente, todavía faltan varias décadas para que sea una realidad** [3]. Si se lograra construir algo así, sería una verdadera revolución para todos los sectores económicos, aunque también introduciría numerosos retos éticos, sociales y jurídicos para la humanidad en su conjunto [3, 4]. Por su importancia, algunos autores denominan a este futuro hito histórico «singularidad tecnológica» *(technological singularity)* [3]. Es esencial destacar que **numerosas aplicaciones teóricas de la IA** (por ejemplo, el análisis automático –con fiabilidad absoluta– de un extenso texto jurídico escrito en lenguaje natural) **requieren de las capacidades de una IAG y,** por tanto, **resultan inviables con la tecnología actual.**

El estudio de las técnicas y algoritmos de IA se suele denominar aprendizaje estadístico *(statistical learning)*, **aprendizaje automático** *(machine learning)*, y –con menor frecuencia– **minería de datos** *(data mining)*[6] [1, 2]. Se suelen diferenciar los métodos que emplean un aprendizaje guiado, con datos etiquetados *(supervised learning)* de los que usan un aprendizaje autónomo, sin datos etiquetados *(unsupervised learning)*, existiendo además los de aprendizaje por refuerzo *(reinforcement learning)*[7] [1, 2]. Desde el punto de vista de la Estadística, estos planteamientos pueden considerarse como una evolución de la **estadística**

[1] Por ello, en general, es recomendable que la selección inicial de las variables de un algoritmo sea realizada por expertos humanos.

[2] Por ejemplo, un modelo de IA capaz de predecir con una fiabilidad muy elevada una enfermedad tomando como entrada imágenes radiológicas de un hospital, puede dar resultados sensiblemente peores en otro hospital, debido al ajuste cromático o a otras diferencias de calibración de los aparatos de los dos hospitales.

[3] Por ejemplo, en visión artificial frecuentemente se emplean modelos con cientos de variables de entrada, cuya comprensión detallada resulta inabarcable para un ser humano.

[4] Es el caso de numerosos algoritmos de IA modernos. Por ejemplo, las redes neuronales artificiales *(artificial neural networks)*, inspiradas en el comportamiento de las neuronas biológicas, tienen la capacidad de modelar fenómenos muy complejos, pero los modelos resultantes son difíciles de interpretar para los seres humanos, pudiendo dar lugar al «efecto de caja negra».

[5] De hecho, en diversas organizaciones frecuentemente se llega a la conclusión de que es más sencillo sustituir un sistema basado en algoritmos de IA por una serie de consultas en lenguaje SQL.

[6] La denominación más ajustada a la realidad metodológica seguramente sea *statistical learning* (en el sentido de aprendizaje basado en métodos derivados de la estadística), pero las estrategias de mercadotecnia hicieron más populares los términos *machine learning* y *data mining*.

[7] En este caso se intenta que un agente *software* aprenda a tomar medidas en un entorno para ir maximizando alguna métrica (a la que se suele denominar «recompensa»). Por ejemplo, una de las maneras más eficaces de lograr que un pequeño helicóptero controlado por IA aprenda a realizar ciertas tareas, es dejar que vuele en un entorno físico controlado y establecer métricas de recompensa y castigo, para ir ajustando el aprendizaje.

descriptiva[8] y **predictiva**[9], así como del **modelado de procesos estocásticos**[10]. Por otra parte, muchos algoritmos de IA modernos se clasifican en las ramas de **optimización convexa y no convexa** de las ciencias matemáticas. Además, la IA se inspira en otras disciplinas tales como la teoría de juegos *(game theory)*, la investigación operativa *(operations research)*, la teoría del control *(control theory)* o la teoría de información *(information theory)* [2].

Dado que los algoritmos de IA actuales poseen una capacidad de generalización inferior en varios órdenes de magnitud a la de un ser humano, normalmente, requieren grandes cantidades de datos para su entrenamiento. El área de las ciencias de la computación que estudia el procesamiento masivo de datos, a veces denominado *big data*[11], se suele asociar a la IA, aunque –en realidad– la IA no siempre requiere del uso de tecnologías *big data* (como veremos, a veces se pueden lograr resultados aceptables sin bases de datos gigantescas) y el *big data* tiene otras aplicaciones además de ser fuente de algoritmos de IA (tales como el cálculo de indicadores sencillos sobre fuentes de datos de gran tamaño).

VIABILIDAD TÉCNICA Y ORGANIZATIVA DE LOS PROYECTOS DE IA

La siguiente cita, del año 1957, se atribuye al investigador Herbert A. Simon: «No quiero sorprenderles o alarmarles, pero (...) ahora mismo existen en el mundo máquinas que piensan, aprenden y crean. Además, su capacidad para realizar estas actividades crecerá rápidamente hasta que, en un futuro previsible, el rango de problemas que pueden acometer sea similar al de la mente humana (...)» [2]. Esta afirmación, a pesar de su optimismo temerario, podía tener cierta justificación en la época, dado que algunos problemas sencillos se habían logrado resolver mediante algoritmos innovadores de IA[12], y se podía pensar que se lograrían avances significativos en poco tiempo. Sin embargo, muy pronto se descubrió que los algoritmos que se empleaban eran inviables por su complejidad computacional [5, 6]: la necesidad de potencia de cálculo crecía de manera prohibitiva con el tamaño del problema y la tecnología de la época no permitía su abordaje [2].

Ya en los años 80, con una visión más templada y realista, el transhumanista y experto en robótica Hans Moravec [7] planteó su famosa paradoja: «es relativamente sencillo lograr que las computadoras exhiban las habilidades de un adulto en pruebas de inteligencia o juegos como las damas, pero es difícil o imposible dotarlas de las habilidades perceptivas y motrices de un niño de un año de edad». En gran medida, estas restricciones continúan vigentes en la actualidad, a pesar del enorme crecimiento de la capacidad computacional[13], de la mejora de los algoritmos de IA y de la disponibilidad de grandes bases de datos, que deberían poder proporcionar información a dichos algoritmos. De hecho, algunos de los avances en IA que han tenido mayor difusión mediática (tales como el sistema AlphaZero [8] de Google; o el sistema TorchCraft [9] de Facebook), siguen siendo ejemplos de sistemas automáticos expertos en juegos similares a las damas (ajedrez, *shogi, go* o Starcraft), aunque con tableros más grandes y piezas que pueden realizar acciones mucho más complejas. Por otra parte, aunque en los últimos 30 años se han producido avances significativos en visión artificial, audición artificial y motricidad automática, los sistemas de IA actuales siguen teniendo limitaciones importantes[14].

En 2016, una de las figuras más célebres de la IA aplicada, Andrew Yan-Tak

[8] Con métodos de *clustering* o árboles CART *(Classification And Regression Trees)*, entre otros.

[9] Tales como los algoritmos de regresión logística clásica o penalizada, *Support Vector Machines* (SVMs), entre otros.

[10] Siendo especialmente relevante la familia de modelos de Márkov: *Markov chain, hidden Markov model (HMM), Markov decision process (MDP)*.

[11] Cabe destacar que el término *big data* no tiene una única definición, y es un término más comercial que técnico.

[12] De hecho, muchos algoritmos de IA fueron desarrollados a nivel teórico en esta época.

[13] Algunos algoritmos de IA comenzaron a ser viables en la práctica por el desarrollo del *hardware* en las tres últimas décadas. Por ejemplo, el uso de nuevas CPUs *(central processing units)*, GPUs *(graphic processing units)* y ASICs *(application-specific integrated circuits)* ha permitido un avance importante en las aplicaciones basadas en redes neuronales artificiales. El desarrollo de circuitos con capacidad de cristalización cuántica *(quantum annealing)* podría permitir progresos todavía mayores.

[14] Por ejemplo, un sistema de visión artificial puede ser engañado introduciendo píxels con ruido en ciertas coordenadas de la imagen.

Ng, declaró que «casi todo lo que una persona normal puede hacer en menos de 1 segundo (de pensamiento), se puede automatizar con la IA» [10]. Esto sería posible suponiendo que se cumplieran además una serie de premisas, que veremos a continuación. Es importante destacar que, para algunos problemas, su cumplimiento es imposible; o resulta viable únicamente para empresas líderes en IA a nivel mundial.

La viabilidad técnica y organizativa de una determinada solución de IA en la práctica requiere, al menos, el cumplimiento de los siguientes criterios: **(i) datos representativos y suficientes del fenómeno en estudio; (ii) el impacto asociado a los errores del sistema de IA debe ser asumible para la organización que lo emplee.** Asimismo, en menor medida, son relevantes estos requisitos: **(iii) es necesario un equipo humano con conocimiento suficiente sobre IA y sobre el dominio del problema concreto que se pretende resolver; (iv) se debe disponer de recursos computacionales suficientes.**

Criterios fundamentales: datos representativos y suficientes; impacto asumible de error de uso de sistema de IA

¿Qué quiere decir que los **datos** sean **representativos**? Podemos disponer de muchos datos que no sean representativos del fenómeno en estudio. Un ejemplo clásico en este sentido fue una encuesta electoral realizada por la revista estadounidense *The Literary Digest* en el año 1936 [11]. Dicha revista logró recabar encuestas de más de 2 millones de votantes –un tamaño muestral impresionante para la época–, pero fracasó estrepitosamente en la predicción del resultado electoral, al encuestar predominantemente a votantes del partido republicano, ignorando a gran parte del electorado, que era votante del partido demócrata. En esta misma época surgió la conocida empresa de encuestas Gallup [11], que predijo con éxito el resultado de las elecciones realizando muchas menos encuestas, pero más representativas del electorado estadounidense en su conjunto.

¿Se pueden conseguir datos representativos para cualquier problema? No siempre. Existen fenómenos con un determinado comportamiento durante un período temporal, que después se modifica, por características intrínsecas, extrínsecas o una combinación de ambas[15].

Generalmente, el entrenamiento de un sistema de IA con datos que han perdido su vigencia dará lugar a modelos sesgados y erróneos. De hecho, si la variabilidad temporal de las características del fenómeno en estudio es frecuente y muy elevada, el fenómeno no es predecible de manera fiable[16]. Estas situaciones se pueden identificar teniendo en cuenta las características del problema o detectar mediante técnicas tales como los indicios de fractalidad [12].

¿Qué quiere decir que los **datos** sean **suficientes**? La mayor parte de los algoritmos de IA requieren muchísimos más datos que un ser humano[17], para llegar a conclusiones similares o peores[18]. Estos datos deben ser estructurados y estar adecuadamente depurados. En general, la obtención de datos de calidad tiene un coste económico y en recursos de otra índole. Incluso, en ocasiones, es imposibles conseguir datos suficientes, bien por restricciones jurídicas, o bien porque no existen apenas ejemplos del fenómeno en estudio[19]. **Es por ello que los sistemas actuales de IA se están encontrando con una limitación fundamental en su desarrollo y este es uno de los motivos de la encarnizada lucha por los datos de las grandes empresas de IA** (Google, Facebook, Amazon, Netflix, etc) **y de algunos actores estatales.**

Además de la disponibilidad de datos representativos y suficientes, el **impacto de errores de uso del sistema de IA debe ser asumible para la organización.** ¿Qué es el impacto? No es lo mismo que el error. Una probabilidad de error muy baja (p.ej. 0,1% de resultados incorrectos), pero con

[15] Supongamos que tratamos de construir un sistema que predice la afluencia de pacientes a los servicios de urgencias de una serie de hospitales de una ciudad. El fenómeno en estudio cambiará invalidando los datos históricos si, en un determinado momento, se inauguran hospitales nuevos con servicios de urgencias, o se alteran sensiblemente las características de la población usuaria, p.ej. si se producen inmigraciones o emigraciones masivas.

[16] Por ejemplo, la evolución de la cotización de ciertos activos en algunos mercados financieros presenta estas características.

[17] Generalmente, entre 2 y 5 órdenes de magnitud.

[18] Aunque existen iniciativas que intentan reducir las necesidades de datos para determinados escenarios. Por ejemplo, en visión artificial se han desarrollado las redes neuronales artificiales convolucionales *(convolutional networks)* y de cápsulas *(capsule networks)*.

[19] Por ejemplo, uno de los grandes problemas del estudio de las enfermedades raras es precisamente la falta de datos, debida a su reducida prevalencia.

	IMPACTO BAJO DE UN ERROR DEL SISTEMA DE IA	IMPACTO ALTO DE UN ERROR DEL SISTEMA DE IA
DATOS REPRESENTATIVOS Y SUFICIENTES	Proyectos viables en producción. *[cuadrante II]*	Proyectos con potencial de impactos muy negativos de baja probabilidad[22] en producción. Para valorar su uso en producción se deben paliar o eliminar los impactos de los errores. *[cuadrante I]*
DATOS NO REPRESENTATIVOS Y/O INSUFICIENTES	Proyectos viables únicamente como pilotos o pruebas de concepto. Para valorar su uso en producción es esencial conseguir datos suficientes y representativos. *[cuadrante III]*	Proyectos con elevada probabilidad de consecuencias catastróficas en producción. No deberían emprenderse. *[cuadrante IV]*

Tabla 1. Evaluación de viabilidad técnica y organizativa de proyectos de IA para sistemas en producción.

Nota: existen fenómenos para los que resulta difícil o incluso imposible obtener datos suficientes y representativos. Un ejemplo es la evolución de la cotización de ciertos activos en algunos mercados financieros con una variabilidad muy elevada. Otro ejemplo son las enfermedades de muy baja prevalencia, para las que existen pocos casos de estudio.

impactos (consecuencias) potencialmente catastróficos asociados a este error (p.ej. fallecimiento de un paciente; fallo en el sistema de control de una central nuclear) puede ser inasumible para una organización. Por ello, **a la hora de evaluar modelos de IA, deberían emplearse métricas de impactos y no solamente métricas de error.** A esto se debe añadir que el impacto derivado del uso de un sistema de IA puede no ser lineal, p.ej.: un error de 10 unidades puede dar lugar a una pérdida económica de 100€; un error de 20 unidades, puede resultar en una pérdida de 1.000€ y un error de 30 unidades, de 10.000€. Por ello, es fundamental estimar los impactos de manera exhaustiva antes de plantear el uso de un sistema de IA en producción. **La evaluación de los errores e impactos de un modelo de IA debe realizarse sobre varios conjuntos de datos representativos, que deben ser distintos de los datos empleados para su entrenamiento[20].**

Cabe destacar que este marco de razonamiento es muy poco habitual en la literatura académica. Generalmente, en las publicaciones científicas en las que se evalúan modelos de IA, los autores se limitan a mencionar las **métricas de error de sus modelos[21]**, pero omiten los posibles **impactos de los errores**, en parte, porque esto último depende de las características de la organización que decida emplear el modelo de IA en producción.

Partiendo de estos criterios fundamentales, se puede construir el cuadrante de toma de decisión de la **Tabla 1.**

[20] Para ello existen diversas técnicas tales como la separación manual en conjuntos de entrenamiento y validación, la validación cruzada *(K-fold cross validation)* o la validación de ventana deslizante con ordenación temporal.

[21] Asimismo, en los trabajos académicos frecuentemente se emplean métricas de error absoluto, que dan el mismo peso al error por exceso que al error por defecto. Esto no suele corresponderse con los impactos asociados a estos errores. Por ejemplo, un error por exceso en la dotación de personal de un servicio de urgencias de un hospital puede interpretarse como un uso ineficiente de los recursos, pero un error por defecto puede dar lugar a una crisis asistencial.

[22] El autor N. N. Taleb denominó este escenario «cisne negro» o *black swan*, por similitud con una situación aparentemente imposible, pero que sí puede darse en la realidad (durante siglos en Europa se creyó que únicamente existían cisnes blancos; pero en el s. XVIII los europeos descubrieron que existían cisnes negros, en Australia).

A continuación, analizaremos algunos ejemplos de este marco de razonamiento.

EJEMPLO 1: un sistema desarrollado por Google para el diagnóstico automático de la retinopatía diabética basado en una base de datos de más de 128.000 retinografías etiquetadas por oftalmólogos [13]. Se valora el uso de este sistema en la India, cuya población presenta una prevalencia elevada de la enfermedad y donde existe un grave déficit de oftalmólogos. Los datos son representativos y suficientes (el tamaño muestral es muy amplio) y al modelo predictivo tiene un error mínimo. El impacto de un error puede ser la ceguera de algún paciente, pero el impacto del no uso del sistema es mucho mayor para la población de la India, debido a su déficit de oftalmólogos. Por ello, sería un sistema de cuadrante II, cuyo uso podría ser viable en producción.

EJEMPLO 2: una gran fabricante de automóviles se plantea la introducción de un nuevo modelo de coche con un sistema de IA para una conducción completamente autónoma (lo que se denomina «nivel de autonomía 5» o «autonomía total» [14]), en todo tipo de condiciones meteorológicas; en cualquier escenario de tráfico y en cualquier lugar del mundo. En este caso, aun teniendo una base de datos enorme para el entrenamiento del algoritmo, nos situaríamos en un escenario de datos posiblemente insuficientes y seguramente no representativos de toda la casuística posible. Asimismo, el impacto del error podría ser muy elevado: podrían morir decenas de personas en accidentes causados por vehículos autónomos. Este sería un proyecto de cuadrante IV, que no debería emprenderse. Una posible solución sería acotar el alcance del sistema de IA, restringiéndolo, por ejemplo, a camiones autónomos que realizaran únicamente rutas muy concretas por carriles especiales con sistemas que les ayudaran a orientarse (p.ej. con radiobalizas y sensores de ultrasonidos). Con estos condicionantes, se podría reubicar el problema en el cuadrante I o incluso en el cuadrante II (p.ej. introduciendo sistemas de frenado automático ante situaciones imprevistas).

Criterios secundarios: equipo humano y recursos computacionales suficientes

Además de los criterios fundamentales enunciados en el apartado anterior (datos representativos y suficientes; riesgo de uso de IA asumible para la organización), habría que mencionar además una serie de criterios secundarios, para los que existen soluciones, si se dispone de recursos financieros, y si el tiempo de ejecución del proyecto es suficientemente dilatado.

En primer lugar, **es necesario un equipo humano, interno o externo, con conocimiento suficiente en IA y en el dominio del problema.** Si bien es cierto que existen numerosas áreas de las Tecnologías de la Información y de las Comunicaciones (TIC) que se han estandarizado y cuyo coste ha disminuido de manera importante en los últimos años, una tecnología nueva como la IA generalmente requiere personal experto en la materia. Cabe destacar que, además, tanto el coste como el tiempo de ejecución del proyecto se reducirán considerablemente si existe un **grado de madurez suficiente de la IA en el dominio del problema** y el equipo humano está familiarizado con problemas similares. Por ejemplo, la IA posee una madurez elevada en aplicaciones de análisis de sentimientos en publicaciones en redes sociales en inglés[23], pero su grado de desarrollo es mucho más reducido en aplicaciones de procesamiento de texto libre sobre temas jurídicos en español.

Asimismo, **se debe disponer de recursos computacionales suficientes para el entrenamiento de los algoritmos y para la validación de los modelos.** La complejidad computacional de algunos algoritmos de IA actuales sigue siendo muy elevada [5, 6] y, a pesar de los avances en *hardware,* los problemas complejos de IA en ocasiones requieren de mayor potencia computacional de la instalada en el centro de proceso de datos (CPD) de una organización. En estos casos se puede acudir a soluciones *cloud,* que permiten la absorción de demanda de potencia computacional necesaria para el entrenamiento de algoritmos de IA, aunque su coste siempre debe ser comparado con el de la adquisición de *hardware* de CPD[24].

[23] Existiendo numerosas aplicaciones de código abierto que se podrían reutilizar para aplicaciones similares.

[24] Actualmente, para aplicaciones de IA suele ser más rentable adquirir *hardware* para el CPD que alquilarlo en *cloud,* si se le da un uso continuado, aunque en este tipo de estudios es importante valorar los costes totales de uso (costes de *hardware* y *software,* costes operativos y costes a largo plazo).

VIABILIDAD JURÍDICA DE LOS PROYECTOS DE IA EN LAS ADMINISTRACIONES PÚBLICAS

A continuación, realizamos un breve análisis jurídico del uso de sistemas de IA en las Administraciones Públicas, fundamentado en la legislación básica conformada por la Ley 39/2015, de 1 de octubre, del Procedimiento Administrativo Común de las Administraciones Públicas (LPAC) y la Ley 40/2015, de 1 de octubre, de Régimen Jurídico del Sector Público (LRJSP). Dado el artículo 156 de la LRJSP, serían de aplicación los Reales Decretos 3/2010 y 4/2010, de los Esquemas Nacionales de Seguridad e Interoperabilidad (ENS y ENI), así como sus normas e instrucciones técnicas de desarrollo, aunque de momento no poseen provisiones específicas sobre IA. Además, si existe tratamiento de datos de personas físicas, serían aplicables tanto el Reglamento (UE) 2016/679 (RGPD), como la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales (LOPDP). Dejando a un lado la normativa sectorial y la legislación en materia de transparencia y reutilización del sector público[25], **identificamos tres escenarios jurídicos, que se describen a continuación en los siguientes apartados: (i) sistemas de IA para la actuación administrativa automatizada, sin intervención humana; (ii) sistemas de soporte a la decisión humana basados en IA; (iii) sistemas de información a la ciudadanía.** En todos los casos, sería esencial asegurarse de que el sistema de IA no empleara variables de tal forma que pudiera comprometer libertades o derechos de los administrados de manera antijurídica.

Sistemas de IA para la actuación administrativa automatizada, sin intervención humana

En una publicación reciente [15], el catedrático de derecho administrativo J. Valero Torrijos identifica diversas lagunas regulatorias y plantea recomendaciones sobre **sistemas de IA capaces de decisiones administrativas sin supervisión humana.** A continuación, se realiza un resumen de algunos de sus puntos más destacados.

En primer lugar, sería de aplicación el artículo 41 de la LRJSP, referente a las **actuaciones administrativas automatizadas.** Dicho artículo contempla una serie de garantías, tanto para **personas físicas** como para **personas jurídicas,** entre las que cabe destacar la necesidad de definir previamente «las especificaciones, programación, mantenimiento, supervisión y control de calidad y, en su caso, auditoría del sistema de información y de su código fuente». En opinión de J. Valero Torrijos, tanto estos criterios, como la propia decisión de puesta en funcionamiento un sistema de decisión automatizada basada en IA, deberían recogerse en un acto administrativo objeto de publicación en los términos previstos en el artículo 45 de la LPAC [15]. Asimismo, si los afectados potenciales fueran **personas físicas** (no así las jurídicas), sería de especial relevancia el artículo 22 del RGPD, referente a las decisiones individuales automatizadas, y por el que sería necesaria una regulación y el establecimiento de garantías para el afectado, para que este «pudiera acudir a las vías administrativas y judiciales de tutela de los derechos fundamentales» [15]. Además, las exigencias y garantías para personas físicas y jurídicas «deberían someterse a evaluación periódica, para garantizar que se siguen ajustando a las premisas iniciales» [15].

En materia de **responsabilidad patrimonial,** en el supuesto de puesta en marcha de sistemas automatizados de decisión, en general, la Administración asumiría los daños que pudieran derivarse de su funcionamiento, si bien existen diversos supuestos de cierta complejidad –dependiendo de los actores y fuentes de información–, que requieren análisis más pormenorizados. Por ejemplo, «si el algoritmo hubiese sido desarrollado por una entidad privada, la Administración solo podría considerarse responsable en aquellos supuestos en que el daño fuese consecuencia inmediata y directa de una orden suya al amparo de lo dispuesto en el artículo 196 de la Ley 9/2017, de 8 de noviembre, de Contratos del Sector Público» [15].

Teniendo en cuenta el estado actual de la tecnología, los sistemas de decisión administrativa completamente automática basados en IA pueden implicar riesgos relevantes, tanto para la Administración como para potenciales afectados. Por ello, **para las actuaciones administrativas automatizadas normalmente es preferible el uso de código fuente completamente predecible y de actuación transparente, no basado en IA.**

[25] Existen diversas implicaciones para los sistemas basados en IA en el contexto de la Ley 19/2013, de 9 de diciembre, de Transparencia, Acceso a la Información Pública y Buen Gobierno; y de las Leyes 37/2007 y 18/2015 relativas a la Reutilización de Información del Sector Público (RISP). Sin embargo, no creemos que el cumplimiento de dichas leyes comprometa la viabilidad de los proyectos de IA en la mayor parte de los escenarios, aunque pueda incrementar sus costes de ejecución. Por ello, las excluimos intencionadamente de este análisis.

Sistemas de soporte a la decisión humana basados en IA

Por otra parte, los **sistemas de soporte a la decisión basados en IA** –en los que la decisión final es aprobada por un operador humano–, **implican riesgos jurídicos menores y son más viables dada la madurez actual de la IA.** Algunos ejemplos de sistemas de soporte a la decisión basados en IA para las Administraciones Públicas pueden ser los buscadores semánticos especializados, los modelos que agrupen información similar, los sistemas de asignación de metadatos (o de codificación), los generadores de resúmenes de textos, los modelos predictivos de ciertas situaciones puntuales o en serie temporal, los sistemas de detección de datos sensibles dentro de las organizaciones *(Data Loss Prevention)*, algunos sistemas SIEM *(Security Information and Event Management)*, así como los modelos de detección de desviaciones de la norma (que, por ejemplo, podrían dar lugar a actuaciones inspectoras).

En estos casos únicamente se estaría empleando una nueva herramienta *software,* que podría agilizar las actuaciones de la administración, pero no implicaría la aplicación del artículo 41 de la LRJSP, ni exigiría un acto administrativo al efecto. Sin embargo, desde el punto de vista organizativo, sería necesaria la cautela para que el sistema no se convirtiera en uno de actuación administrativa automatizada *de facto,* es decir, uno en el que se confiara ciegamente y el operador humano actuara con un autómata. Por otra parte, si se tratan datos de personas físicas, seguiría siendo necesaria la observancia de la LOPDP y del RGPD, en especial, los criterios de licitud del tratamiento de datos del artículo 6 del RGPD.

Asimismo, sería fundamental el seguimiento de las buenas prácticas de **documentación de las premisas de los algoritmos de IA, la metodología de construcción de los modelos, así como su evaluación periódica** por parte de expertos humanos. Para facilitar estas tareas, **es preferible el uso de algoritmos sencillos, con un número reducido de variables de entrada** seleccionadas por personal experto en la materia. En este sentido, son de gran utilidad los nomogramas [16, 17], las representaciones gráficas de árboles de decisión [18] o las gráficas de probabilidad de los algoritmos bayesianos [19].

Sistemas de información a la ciudadanía

Otra **área de aplicación de IA bastante amplia son los sistemas automáticos de información a la ciudadanía,** sin intervención humana, que incluye numerosos ejemplos tales como *chatbots,* asistentes virtuales, o algunos tipos de sistemas de recomendación *(recommender systems)*.

En estas aplicaciones de la IA, la regulación es difusa, dado que no dan lugar a una actuación administrativa automatizada, ni son realmente sistemas de soporte a la decisión para el personal al servicio de la Administración; y, generalmente, se parte de la debatible asunción de que no proporcionan información jurídicamente vinculante. Sin embargo, **es esencial que estos sistemas se nutran de datos actualizados y fiables, que den lugar a información veraz, y que estén sometidos a una auditoría de calidad continua por parte de actores humanos.**

CONCLUSIONES

La IA será un factor relevante para la transformación digital de la economía de España y de su sector público y, precisamente por ello, **en los próximos años será esencial que los recursos y esfuerzos se vuelquen en proyectos de IA con elevadas probabilidades de éxito, que aporten valor real a la ciudadanía, al tejido empresarial y a la Administración.** Para ayudar a identificar estos proyectos, en este artículo se han analizado las definiciones intuitivas y formales de la Inteligencia Artificial, así como sus limitaciones actuales. A continuación, se ha formalizado una **metodología de evaluación de viabilidad técnica y organizativa de proyectos de IA;** y se han esbozado **retos y oportunidades de su uso por parte de la Administración en el marco jurídico actual.** ✱

Bibliografía

[1] I. Goodfellow, Y. Bengio, A. Courville, Deep learning: MIT press, 2016.

[2] S. J. Russell and P. Norvig, Artificial intelligence: a modern approach: Malaysia; Pearson Education Limited, 2016.

[3] A. H. Eden, J. H. Moor, J. H. Søraker, E. Steinhart, Singularity Hypotheses: Springer, 2015.

[4] Y. N. Harari, 21 Lessons for the 21st Century: Jonathan Cape, 2018.

[5] T.-S. Lim, W.-Y. Loh, Y.-S. Shih, "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms", Machine learning, vol. 40, pp. 203-228, 2000.

[6] (2018). Computational complexity of machine learning algorithms. URL: https://www.thekerneltrip.com/machine/learning/computational-complexity-learning-algorithms/

[7] K. Agrawal, "To study the phenomenon of the Moravec's Paradox", arXiv preprint arXiv:1012.3148, 2010.

[8] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, et al., "Mastering chess and shogi by self-play with a general reinforcement learning algorithm", arXiv preprint arXiv:1712.01815, 2017.

[9] G. Synnaeve, N. Nardelli, A. Auvolat, S. Chintala, T. Lacroix, Z. Lin, et al., "Torchcraft: a library for machine learning research on real-time strategy games", arXiv preprint arXiv:1611.00625, 2016.

[10] A. Y.-T. Ng. (2016). Pretty much anything that a normal person can do in <1 sec, we can now automate with AI. URL: https://twitter.com/andrewyng/status/788548053745569792

[11] P. Squire, "Why the 1936 Literary Digest poll failed", Public Opinion Quarterly, vol. 52, pp. 125-133, 1988.

[12] T. Di Matteo, "Multi-scaling in finance", Quantitative finance, vol. 7, pp. 21-36, 2007.

[13] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs", JAMA, vol. 316, pp. 2402-2410, 2016.

[14] J. Fleetwood, "Public health, ethics, and autonomous vehicles", American journal of public health, vol. 107, pp. 532-537, 2017.

[15] J. Valero Torrijos, "Las garantías jurídicas de la inteligencia artificial en la actividad administrativa desde la perspectiva de la buena administración.", Revista Catalana de Dret Públic, 2019.

[16] A. Zlotnik. (2015, 2015-12-15). Nomograms in Stata. URL: http://www.zlotnik.net/stata/nomograms

[17] A. Zlotnik and V. Abraira, "A general-purpose nomogram generator for predictive logistic regression models.", Stata Journal, vol. Volume 15, Number 2, 2015.

[18] D. Steinberg and P. Colla, "CART: classification and regression trees", The top ten algorithms in data mining, vol. 9, p. 179, 2009.

[19] B.-T. Zhang, S.-h. Choi, and Y.-w. Jang. (2019). 4190.408 Artificial Intelligence (2019 Spring). Probabilistic Reasoning 3 (Ch 14. Exact Inference in BNs). URL: https://bi.snu.ac.kr/~scai/Courses/4ai19s/slides/chapter14b.pdf