Retos y regulación de la Inteligencia Artificial: la toma de decisiones en los asuntos públicos y la administración de justicia

Alessandro Mantelero (Politecnico di Torino)

La inteligencia artificial (IA) ya forma parte de nuestra vida cotidiana y se utiliza cada vez más para dar forma a nuestra sociedad digital impulsada por los datos. Sin embargo, aún tiene algunas limitaciones, que pueden afectar negativamente a las personas y a la sociedad. En este sentido, las leyes que pretenden regular la IA deben abordar tres retos principales: los sesgos, la falta de transparencia y la naturaleza propietaria de las soluciones de IA.

Para abordar estas cuestiones cada vez más decisivas y sus repercusiones en los derechos y las libertades fundamentales, se han propuesto varias iniciativas. Una contribución eficaz al debate en este ámbito puede resultar de soluciones centradas en la integración de los derechos humanos y sus principios en el diseño de la IA.

Desde esta perspectiva, este capítulo se centra en dos ámbitos fundamentales de aplicación de la IA: la toma de decisiones en asuntos públicos y la justicia digital. En estos ámbitos, se examinarán los principios rectores y los límites que debe respetar el uso de la IA para no vulnerar los derechos fundamentales.

En este análisis también se destaca el papel de la nueva normativa europea sobre la inteligencia artificial, centrada en los riesgos y con especial énfasis en el impacto sobre los derechos fundamentales.

PALABRAS CLAVE : Derechos Fundamentales; Inteligencia Artificial; Democracia Electrónica; Justicia Digital; Regulación

1. Los desafíos de la inteligencia artificial en la sociedades digitales

La Inteligencia Artificial (IA) ya forma parte de nuestra vida cotidiana y se utiliza cada vez más para modelar nuestra sociedad digital basada en datos. Se utiliza para moderar el debate público, modelar el entorno social y apoyar a los responsables humanos de la toma de decisiones en diversos ámbitos, incluida la justicia. La IA es, por lo tanto, un componente de decisiones que afectan a individuos y grupos, contribuyendo a dar forma a nuestras comunidades y vidas.

Para enmarcar correctamente este análisis, es importante tener en cuenta la diferencia entre inteligencia natural e inteligencia artificial, más aún en lo que respecta a la IA Generativa (GAI) con sus recientes aplicaciones basadas en LLMs (large language models), que cada vez se utilizan más para interactuar con los ciudadanos por parte de diversas aplicaciones de la administración pública, así como en mayor medida con la futura IA de Propósito General (GPAI).

A este respecto, cabe señalar que la IA no es más que una forma matemática y basada en datos de procesamiento de la información. La IA no es capaz de pensar, elaborar conceptos o desarrollar teorías: la IA se limita a adoptar un enfoque de reconocimiento de trayectorias para ordenar enormes cantidades de datos e inferir nuevas informaciones y correlaciones.

La dependencia de los datos es a la vez la fuerza y la debilidad de estos sistemas en términos de impacto potencial sobre los derechos fundamentales y la democracia. Datos de escasa calidad producen resultados poco fiables y difunden información sesgada.[1] Además, la 'datificación' ofrece a menudo una representación parcial de la realidad que no incluye a los grupos minoritarios e infrarrepresentados,[2] los conjuntos de datos increíblemente grandes y las complejas soluciones de IA pueden resultar oscuros tanto para los responsables humanos de la toma de decisiones como para las personas o grupos perjudicados. Todo esto afecta a la plena aplicación de los principios que sustentan una gobernanza responsable y transparente.

El resultado de estas limitaciones técnicas y estructurales puede resumirse en tres categorías principales: sesgos, opacidad y propriedad.

En cuanto a los sesgos, el diseño y desarrollo de herramientas de IA puede verse afectado por diferentes sesgos que, en muchos casos, difieren de los sesgos humanos.[3] No sólo afectan a la calidad de los datos (por ejemplo, el sesgo de selección), sino también a las metodologías adoptadas (por ejemplo, sesgos en el preprocesamiento, la limpieza de datos y las metodologías de encuesta y de medición),[4] la naturaleza de la operación de procesamiento de la IA (por ejemplo, las llamadas 'alucinaciones' en la GAI),[5] el objetivo de la investigación (por ejemplo, el sesgo histórico en conjuntos de datos preexistentes y

[1] Véase European Union Agency for Fundamental Rights (2019).
[2] Véanse P. Agre (1994) y M. Hildebrandt (2019).
[3] Véanse R. Caruana et al. (2015) y K. Eykholt et al. (2018).
[4] Véase M. Veale y R. Binns (2017).
[5] En la IA, una alucinación es una respuesta segura de una IA que no está justificada por sus datos de entrenamiento, produciendo información y resultados erróneos.

la infrarrepresentación o sobrerrepresentación de determinados grupos en nuevos conjuntos de datos) y la actitud psicológica de los científicos de datos (sesgo de confirmación).[6]

Esta breve lista de posibles sesgos también revela el componente humano de las soluciones de IA, a menudo subestimado en una comparación engañosa entre humanos y máquinas. Esta dicotomía subestima el papel de la intervención humana en el procesamiento de datos de IA[7] y no considera adecuadamente la transposición, intencionada o no, de las visiones de los desarrolladores en cuanto a los valores de referencia de los modelos de IA.[8]

Con referencia a la opacidad, concierne tanto a las herramientas de IA utilizadas como a la forma en que repercuten en los individuos, cuyas situaciones se analizan y representan a través de ellas. No sólo se desconoce el funcionamiento real y el tratamiento de la información de algunas aplicaciones de IA,[9] incluso para los científicos de datos, sino que los individuos no suelen ser conscientes de que se les agrupa dinámicamente sobre la base de correlaciones e inferencias invisibles, sin poder conocer la identidad de los demás miembros del grupo. Por lo tanto, la opacidad tiene dos consecuencias diferentes: en primer lugar, los científicos de datos no pueden justificar claramente las decisiones específicas sugeridas por la IA y, en segundo lugar, las personas son examinadas pasivamente por la IA sin tener un papel significativo o efectivo en el diseño de la IA ni la oportunidad de expresar sus intereses colectivos.[10]

Este nivel de opacidad y las limitaciones a la participación democrática en el desarrollo de la IA se ven acentuados por una tercera característica de muchos productos de IA: la propiedad. La naturaleza patentada de los algoritmos utilizados y, en cualquier caso, la capacidad de generar silos de datos utilizados para entrenarlos e implementarlos, muestran como las lógicas de apropiación representan una barrera más para acceder a la arquitectura de estas aplicaciones y la supervisión colectiva.[11]

Por último, en lo que respecta a la relación proveedor/utilizadores principal de sistemas de IA,[12] una cuestión crítica está representada por la diferencia presente en la cadena de

[6] Véase R. Nickerson (1998).
[7] Véanse, por esemplo, P. Tubaro et al. (2020); K. Crawford y V. Joler (2018); N.N. Loideain y R. Adams (2020).
[8] Véase S.M. West et al. (2019).
[9] Véanse A.D. Selbst (2017); J. Burrell (2016); R. Brauneis y E.P. Goodman (2018).
[10] Véanse C.B. Graber (2020) y A. Mantelero (2016).
[11] Véanse F. Pasquale (2015: 193).
[12] Aquí y en el texto siguiente se distingue entre proveedor de IA, usuario principal y usuario final. En el caso, por ejemplo, de una solución de IA utilizada por una administración pública para dar respuestas

uso de la IA en términos de (i) competencias, (ii) acceso a algoritmos y datos de entrenamiento, y (iii) conocimiento contextual. Mientras que los proveedores de IA dominan los dos primeros elementos, el último está principalmente en manos del utilizador primario, utilizando este término para identificar a las entidades (empresas u organismos públicos) que utilizan herramientas de IA en un contexto específico y para su propio fin concreto (por ejemplo, una municipalidad que utiliza un chatbot para atender a los ciudadanos que hacen preguntas sobre algunos temas).

Esta diferencia en el papel de los actores implicados no es nueva y ya está contemplada en el RGPD en lo que respecta a la gestión de riesgos, es decir, la Evaluación de Impacto relativa a la Protección de Datos (EIPD), donde una EIPD general del proveedor de un servicio puede complementarse con una EIPD relativa al uso específico del propio servicio realizado para el usuario principal. En cuanto al IA, el punto clave es, por un lado, la dificultad de los proveedores para prever todas las posibles aplicaciones contextuales de una herramienta tecnológica determinada (por ejemplo, los mismos sistemas de videovigilancia inteligente pueden utilizarse en espacios públicos, en zonas con un alto índice de delincuencia, en escuelas, con diferentes niveles de riesgo en términos de impacto sobre los derechos fundamentales) y, por otro, las dificultades del usuario principal (deployer en el AI Act) para acceder a productos/servicios que son en gran medida arquitecturas cerradas, cuando no oscuras.

A este respecto, en consonancia con los principios cardinales del derecho de responsabilidad civil basados en la prevención de riesgos y la correspondiente atribución del deber de diligencia, la gestión del riesgo debe distribuirse de forma proporcional entre los proveedores y los usuarios principal en función del riesgo efectivo introducido por cada uno en la sociedad y del poder efectivo para gestionar el riesgo, tal como se acepta generalmente en la teoría jurídica del riesgo.

Así, unos riesgos se refieren al desarrollo del sistema de IA y otros a su utilización en un escenario concreto. Para los primeros, el proveedor está en la mejor posición para gestionarlos, mientras que el usuarios principal puede concurrir, bajo ciertas condiciones, en la gestión de los segundos. En este contexto, el posible papel desempeñado por el usuarios principal se basa en que es necesaria una gestión contextual de los riesgos debido al contexto específico de uso del sistema, es decir riesgos que no pueden ser previstos o

automáticas en línea a los ciudadanos, el proveedor será la empresa que suministre el software, la administración será el usuario principal y los ciudadanos los usuarios finales. Esta distinción de funciones se introdujo en la versión final del AI Act, que habla de provider, deployer y user.

gestionados adecuadamente a nivel general por el proveedor, por ejemplo, vulnerabilidades específicas de las personas afectadas. La segunda condición para que el usuarios principal lleve a cabo la gestión de riesgos es la viabilidad, es decir, que el sistema sea suficientemente accesible y "personalizable" por el implantador.

Esta línea de pensamiento se plasma ahora en el AI Act, que establece obligaciones específicas para los usuarios principal de IA en relación con la realización de una evaluación de impacto sobre los derechos fundamentales en caso de uso de sistemas de IA de alto riesgo.

Estos tres limites estructurales – sesgos, opacidad y propriedad – y la cuarta cuestión organizativa – relación proveedor/ usuarios principal –, que aquí se han analizado brevemente, tienen un impacto directo en los retos de la IA y su aceptación social en tareas de supervisión y gobierno de las actividades humanas (por ejemplo, las ciudades inteligentes), la oferta de servicios personalizados (por ejemplo, la medicina predictiva o la información en línea a los ciudadanos) y, más en general, el apoyo a los seres humanos en el proceso de toma de decisiones.

Las cuestiones que rodean a las soluciones basadas en el uso intensivo de datos y su utilización en los procesos de toma de decisiones afectan a una serie de intereses relacionados con diversos derechos humanos/fundamentales. No sólo el riesgo de discriminación es uno de los mayores retos de estas aplicaciones, sino que también son importantes otros derechos y libertades, como el derecho a la integridad de la persona, la educación, la igualdad ante la ley, así como la libertad de circulación, pensamiento, expresión, reunión y libertad en el lugar de trabajo.[13]

Para hacer frente a la creciente preocupación por las posibles repercusiones de la IA en los derechos fundamentales y en las libertades, se han propuesto varias iniciativas a escala local, nacional e internacional, y ONG, centros de investigación y entidades empresariales han elaborado diversas directrices. Varias propuestas se han centrado en la ética, a menudo desdibujando la línea entre derecho y ética, describiendo los derechos fundamentales como valores éticos con su "eticidad" y relativización.

Este énfasis en la dimensión ética puede entrañar el riesgo de extender al ámbito del tratamiento de datos un imperialismo ético cuyos efectos ya son conocidos en biomedicina y ciencias sociales. A este respecto, la experiencia previa en materia de evaluación ética de la investigación científica sugiere que debe tenerse en cuenta la

[13] Véase A. Mantelero (2022).

distinción entre valores éticos y jurídicos, así como las diferencias entre enfoques éticos. Varios documentos que proporcionan directrices sobre la IA se refieren al marco ético de forma bastante amplia e indefinida, sin aclarar (o justificar) el marco ético utilizado.

Las respuestas éticas a la incertidumbre en un entorno tecnológico y social en rápida evolución pueden convertirse paradójicamente en una nueva fuente de ambigüedad. Los valores discrecionales y, en algunos casos, basados en intereses, corren el riesgo de debilitar el marco jurídico o de redefinirlo indirectamente sin seguir un procedimiento adecuado, como exige el proceso normativo.

Sin subestimar el papel de la ética en el desarrollo tecnológico, estas consideraciones sugieren una integración más equilibrada entre el derecho y la ética en la regulación de la IA, basada en el énfasis en el papel de los derechos fundamentales como piedra angular de la futura arquitectura de la regulación de la IA. En esta dirección va el modelo Eu definido con el AI Act.

Desde una perspectiva reguladora, el principal reto consiste todavía en contextualizar el marco de los derechos fundamentales definidos por los instrumentos redactados en una era anterior a la IA. En este contexto, se han propuesto iniciativas normativas en varios países, muchas de ellas refiriéndose explícitamente a todos o algunos derechos fundamentales.[14] Sin embargo, a menudo se trata de declaraciones genéricas sin una contextualización adecuada de los derechos y libertades considerados.

Por otro lado, también un enfoque basado en reglas de principios, aunque es relativamente fácil ponerse de acuerdo sobre una lista general de principios para el desarrollo de la IA,[15] sirve de poco para avanzar en el proceso normativo, ya que los principios generales, como la transparencia o la participación, pueden interpretarse de muy diversas maneras.

Por lo tanto, una contribución eficaz al debate sobre los derechos fundamentales en este ámbito sólo puede provenir de una contextualización adecuada en el escenario de la IA. Esto significa adoptar normas operativas que se centren en cómo integrar los derechos fundamentales de forma contextualizada en el diseño de la IA.

En este contexto, las siguientes secciones examinan dos áreas críticas de la aplicación de la IA: toma de decisiones en asuntos públicos y la justicia digital. Estas dos áreas van considerándose centrándose en los aspectos más afectados por las aplicaciones actuales y futuras de la IA.

[14] Véase, por ejemplo, la Carta de Derechos Digitales.

[15] Véase J. Fjeld et al. (2020); F. Raso et al. (2018). Véase también A. Jobin et al. (2019) y T. Hagendorff (2020).

2. Inteligencia artificial y toma de decisiones en asuntos públicos

El derecho a participar en los asuntos públicos se basa en un concepto amplio de asuntos públicos, que incluye el debate público y el diálogo entre los ciudadanos y sus representantes, con un estrecho vínculo con la libertad de expresión, reunión y asociación. En este sentido, la IA es relevante desde dos perspectivas diferentes: como medio para la participación y como objeto de decisiones participativas.

Considerando la IA como medio, las barreras técnicas y educativas pueden socavar el ejercicio del derecho a participar. Por lo tanto, las herramientas de participación basadas en la IA deben tener en cuenta los riesgos de infrarrepresentación y falta de transparencia en los procesos participativos. Al mismo tiempo, la IA también es objeto de decisiones participativas, ya que incluyen decisiones sobre el desarrollo de la IA en general y su uso en los asuntos públicos.

Las plataformas participativas basadas en la IA (por ejemplo, Consul, Citizenlab o Decidim) pueden contribuir significativamente al proceso democrático, facilitando la interacción ciudadana, la priorización de objetivos y los enfoques colaborativos en la toma de decisiones sobre temas de interés general a diferentes niveles (barrio, municipio, área metropolitana, región, país). Dado que estas plataformas se utilizan en un entorno social y recopilan información, cabe recordar aquí la necesidad de tener en cuenta los aspectos relacionados con la protección de datos, incluida la seguridad de datos.

Sin embargo, surgen otras cuestiones más específicas en relación con las herramientas de la IA para la participación democrática (incluidas las de prevención y lucha contra la corrupción), que se asocian a cuatro áreas principales: transparencia, rendición de cuentas, inclusión y apertura. A este respecto, la transparencia es un requisito para el uso de aplicaciones tecnológicas con fines democráticos y es un principio común a otros ámbitos, como la sanidad. Todavía se trata de una noción basada en el contexto: mientras que en la sanidad la transparencia está estrechamente relacionada con la autodeterminación, aquí adquiere un significado más amplio. En un proceso democrático, la transparencia no es sólo un requisito para la autodeterminación de los ciudadanos con respecto a una herramienta técnica, sino que también es un componente del proceso participativo democrático. La transparencia ya no tiene una dimensión individual, sino que asume una dimensión colectiva como garantía del proceso democrático.

En este contexto, el uso de soluciones basadas en la IA para la democracia electrónica debe ser transparente con respecto a su lógica y funcionamiento (por ejemplo, la selección de contenidos en plataformas participativas), proporcionando información clara, fácilmente accesible, inteligible y actualizada sobre las herramientas de IA utilizadas y su justificación.

Además, la aplicación de esta noción de transparencia también debe tener en cuenta la diversidad de usuarios de estas herramientas, adoptando un enfoque accesible desde las primeras fases del diseño de las aplicaciones de IA. Se trata de garantizar una transparencia efectiva con respecto a los grupos vulnerables y discapacitados, dando un valor añadido a la accesibilidad en este contexto.

La transparencia y la accesibilidad están estrechamente relacionadas con la naturaleza de la arquitectura utilizada para construir sistemas de IA. Por lo tanto, el código abierto y los estándares abiertos pueden contribuir a la supervisión democrática de las aplicaciones de IA más críticas. Sin embargo, hay casos en los que el carácter abierto se ve afectado por limitaciones, debido a la naturaleza de la aplicación específica de IA (por ejemplo, la prevención de la delincuencia). En estos casos, la auditabilidad, así como los sistemas de certificación, desempeñan un papel más importante del que ya tienen en relación con los sistemas de IA en general.

En el contexto de las aplicaciones de la IA para fomentar la participación democrática, también puede desempeñar un papel importante la interoperabilidad, ya que facilita la integración entre diferentes servicios/plataformas para la democracia electrónica y a diferentes niveles geográficos. Este aspecto ya es relevante para la democracia electrónica en general, por lo que debería ampliarse al diseño de sistemas basados en IA.

Otro principio clave es la rendición de cuentas. En este sentido, para rendir cuentas, los proveedores de servicios de IA y las entidades que utilicen soluciones basadas en IA para la democracia electrónica tendrán que adoptar formas de vigilancia de algoritmos que promuevan la rendición de cuentas de todas las partes interesadas pertinentes, evaluando y documentando los impactos previstos sobre las personas y la sociedad en cada fase del ciclo de vida del sistema de IA de forma continua, para garantizar el cumplimiento de los derechos fundamentales. Por eso tiene un papel central la evaluación de impacto sobre los derecho fundamentales y sus accesibilidad.[16]

[16] Este papel está ahora expresamente reconocido en el AI Act. Para profundizar en el tema de la evaluación del impacto sobre los derechos fundamentales, véase A. Mantelero (2022: 45-91).

Al abordar los diferentes aspectos del desarrollo de soluciones de IA para la participación democrática, una primera consideración es que un enfoque democrático es incompatible con un enfoque tecno-determinista. Por lo tanto, las soluciones de IA para los problemas de la sociedad deben ser el resultado de un proceso inclusivo. De ahí que valores jurídicos como la protección de las minorías, el pluralismo y la diversidad deberían ser una consideración necesaria en el desarrollo de estas soluciones.

Desde una perspectiva democrática, la primera pregunta que debemos hacernos es: ¿realmente necesitamos una solución basada en la IA para un problema determinado frente a otras opciones, teniendo en cuenta el impacto potencial de la IA sobre los derechos y las libertades? Si la respuesta a esta pregunta es afirmativa, el siguiente paso es examinar la integración de valores en el desarrollo de la IA.

Las soluciones de IA propuestas deben diseñarse desde una perspectiva orientada a los derechos, garantizando el pleno respeto de los derechos y las libertades fundamentales, incluida la adopción de herramientas y procedimientos de evaluación para este fin. En el caso de aplicaciones de IA con un alto impacto en los derechos fundamentales y las libertades, como los procesos electorales, es necesario evaluar previamente el cumplimiento legal. Además, los sistemas de IA para tareas públicas han de ser auditables y, cuando no lo excluyan intereses predominantes contrapuestos, las auditorías estar a disposición del público.

Otro aspecto importante que debe tenerse en cuenta es la colaboración público-privada que suele caracterizar a los servicios de IA para los ciudadanos, sopesando cuál es la mejor opción entre las soluciones propias y las de terceros, incluidas las múltiples combinaciones de estos dos. A este respecto, cuando las soluciones de IA son desarrolladas total o parcialmente por empresas privadas, la transparencia de los contratos y unas normas claras sobre el acceso y el uso de los datos de los ciudadanos tienen un valor crítico en términos de supervisión democrática.

Las restricciones sobre el acceso y el uso de los datos de los ciudadanos no sólo son pertinentes desde la perspectiva de la protección de datos (principios de minimización de datos y limitación de la finalidad), sino más en general con respecto al volumen de datos generados por una comunidad, que también incluye datos no personales y datos agregados. Esta cuestión debería considerarse como un componente de la democracia en el entorno digital, en el que la dimensión colectiva de los recursos digitales generados por una comunidad tendría que conllevar formas de control y supervisión por parte de

los ciudadanos, al igual que ocurre con los demás recursos de un territorio/comunidad (por ejemplo, el medio ambiente).

Conviene recordar aquí las consideraciones ya expresadas anteriormente sobre la apertura como elemento clave de las herramientas de participación democrática, dado su impacto en el diseño de los sistemas de IA. Además, el diseño, el desarrollo y el uso de estos sistemas también tendría que tener en cuenta la adopción de una estrategia respetuosa con el medio ambiente y sostenible.[17]

Por último, cabe señalar que, aunque el diseño de la IA es un componente clave de estos sistemas, el diseño no es neutral. Los artefactos tecnológicos, incluidos los sistemas de IA, pueden incorporar valores. Estos valores pueden elegirse intencionadamente y, en el contexto de la democracia electrónica, deben basarse en un proceso democrático. Pero también pueden integrarse involuntariamente en las soluciones de IA, debido a la composición cultural, social y de género de los equipos de desarrolladores de IA. Por esta razón, la inclusión tiene aquí un valor añadido, en términos de inclusión y diversidad en el desarrollo de la IA.

Los principios debatidos para la democracia electrónica pueden repetirse por lo que respecta a la buena gobernanza. Es el caso de las ciudades inteligentes y la gestión medioambiental basada en sensores, donde los procesos de toma de decisiones abiertos, transparentes e inclusivos desempeñan un papel central. Del mismo modo, el uso de la IA para supervisar las actividades de las autoridades locales, con fines por ejemplo de lucha contra la corrupción, tiene que basarse en el carácter abierto de las soluciones adoptadas (software de código abierto), la transparencia y la posibilidad de auditoría.

En términos más generales, la IA puede utilizarse en la interacción entre el gobierno y los ciudadanos para automatizar las consultas y solicitudes de información de los ciudadanos. Sin embargo, en estos casos, es importante garantizar el derecho a conocer que estamos interactuando con una máquina y a tener un punto de contacto humano. Además, el acceso a los servicios públicos no ha de depender del suministro de datos innecesarios y no proporcionados a la finalidad.

También debe prestarse especial atención al uso potencial de la IA en la interacción hombre-máquina para aplicar estrategias de nudging. En este caso, debido a la complejidad y opacidad de las soluciones técnicas adoptadas, la IA puede aumentar el papel pasivo de los ciudadanos y afectar negativamente al proceso democrático de toma

[17] En el AI Act se incluyen ahora referencias específicas al medio ambiente y la sostenibilidad, también en relación con los llamados modelos fundacionales.

de decisiones. Por eso, tendería que preferirse un enfoque activo basado en la participación consciente y activa en los objetivos de la comunidad, mejor gestionada por las herramientas de participación de la IA.

Por último, el uso de sistemas de IA en tareas de gobierno plantea cuestiones difíciles sobre la relación entre los responsables humanos y el papel de la IA en el proceso de toma de decisiones. Estas cuestiones son más relevantes en relación con las funciones que tienen un alto impacto en los derechos y libertades individuales, como en el caso de las decisiones jurisdiccionales. Por este motivo, en la sección sucesiva se analizarán las preocupaciones sobre la transparencia (incluida la a posibilidad de explicar) del razonamiento de la IA y la relación entre el uso de la IA y la libertad de los responsables de la toma de decisiones.

3. El uso de la IA en la administración de justicia

El ámbito de la justicia es muy amplio y es demasiado ambicioso analizar todo el espectro de consecuencias del uso del IA en este entorno. Por la variedad de tipos y finalidades de las operaciones en este ámbito y las diversas figuras profesionales y procedimientos implicados, en esta sección se hace una distinción funcional entre dos áreas: (i) las decisiones judiciales y las resoluciones alternativas de litigios (ADR) y (ii) la prevención/predicción de la delincuencia. Antes de analizar y contextualizar los principios clave relativos a estas dos áreas, es necesario formular algunas observaciones generales, que también pueden aplicarse a la actuación de la administración pública en su conjunto.

En primer lugar, cabe señalar que, en comparación con las decisiones humanas, y más concretamente con las decisiones judiciales, la lógica de los sistemas de IA no se asemeja al razonamiento jurídico. En su lugar, se limitan a ejecutar códigos basados en un enfoque matemático/estadístico y centrado en datos masivos.

Además, los porcentajes de error de la IA se aproximan o son inferiores a los del cerebro humano en campos como el etiquetado de imágenes, pero las tareas de toma de decisiones más complicadas presentan porcentajes de error más elevados. Es el caso del razonamiento jurídico en la resolución de problemas y vale también para las más recientes aplicaciones basadas en LLMs. Al mismo tiempo, sesgos y ‘alucinaciones’ de la IA en las decisiones jurídicas tienen un alto impacto en los derechos y la libertad de las personas.

Merece también la pena señalar que la diferencia entre los errores en la toma de decisiones humanas y de las máquinas tiene una consecuencia importante en términos de escala: mientras que el error humano afecta sólo a casos individuales, los fallos de diseño y sesgos de la IA afectan inevitablemente a todas las personas en circunstancias iguales o similares, al aplicarse las herramientas de IA a toda una serie de casos. Esto puede causar discriminación de grupo, afectando negativamente a individuos pertenecientes a diferentes categorías tradicionales y no tradicionales. Por otro lado, las alucinaciones que afectan los LLMs pueden surgir y repartirse en manera incontrolada en la serie de caso y con efectos tan deformantes que la misma certeza de la lógica jurídica no se puede garantizar.

Todavía, por la naturaleza textual de los documentos jurídicos, el procesamiento del lenguaje natural (PLN) y el uso del modelos a gran escala (LLMs) pueden desempeñar un papel importante en las aplicaciones de IA en el ámbito de la justicia. Esto plantea varias cuestiones críticas relacionadas con las soluciones existentes desarrolladas con un enfoque en el mercado anglófono, lo que las hace menos eficaces en un entorno jurídico que utiliza lenguas distintas del inglés.

Además, las decisiones jurídicas se caracterizan a menudo por un razonamiento implícito no expresado, que puede ser incorporado en sistemas expertos, pero es más difícil de entenderse para las herramientas de aprendizaje automático basadas en el lenguaje. Por último, la presencia de principios generales y normas abiertas requiere un conocimiento previo de la interpretación jurídica pertinente y actualizaciones continuas que no pueden derivarse de la minería de textos y necesitan una interpolación entre los conocimientos generales de las teorías jurídicas, las interpretaciones de la jurisprudencia y el caso específico.

Todas estas limitaciones sugieren una adopción cuidadosa y crítica de la IA más aún en el ámbito de la justicia que en otros ámbitos y, por lo que respecta a las resoluciones judiciales y las ADR (resolución alternativa de litigios), recomiendan una distinción entre los casos caracterizados por evaluaciones rutinarias y basadas en hechos y los casos en lo cuales el razonamiento jurídico y la discrecionalidad juegan un papel importante.

En primer lugar, cabe destacar como varios de los productos de IA utilizados en el entorno jurídico no tienen un impacto directo en los procesos de toma de decisiones en los tribunales y tampoco en las resolución alternativa de litigios, sino que facilitan la gestión de contenidos y conocimientos, la gestión organizativa y la medición del

rendimiento. Estas aplicaciones incluyen, por ejemplo, herramientas de categorización de contratos, detección de cláusulas contractuales divergentes o incompatibles, e-discovery, ayuda a la redacción de textos jurídicos, recuperación de disposiciones legales, revisión asistida del cumplimiento. Además, algunas aplicaciones pueden ofrecer funciones básicas de resolución de problemas basadas en preguntas estándar y situaciones normalizadas (por ejemplo, chatbots jurídicos).

Aunque en estos casos la IA tiene un impacto en la práctica jurídica y en los conocimientos jurídicos que plantea diversas cuestiones éticas, las posibles consecuencias adversas para los derechos fundamentales, en la prospectiva trazada por AI Act, son más limitadas. En gran medida, están relacionadas con ineficiencias o defectos de estos sistemas y pueden ser enfrentadas con estrategias adecuadas.

En el caso de la gestión de contenidos y conocimientos, incluida la investigación y el análisis de documentos, estos fallos pueden generar representaciones incompletas o inexactas de hechos o situaciones, pero esto afecta a los metaproductos, los resultados de una herramienta de investigación que deben interpretarse y motivarse adecuadamente cuando se utilizan ante un tribunal. La intervención humana por parte de los utilizadores expertos en las profesiones jurídicas puede razonablemente reducir estos riesgos y, de todas formas, siempre se pueden aplicar las normas de responsabilidad, en el contexto de la responsabilidad por productos defectuosos, como protección ex post.

Además, el sesgo (mala selección de casos, clasificación errónea, etc.) que afecta a las herramientas basadas en texto para el análisis de la legislación, la jurisprudencia y la literatura, puede contrarrestarse mediante una educación y formación adecuadas de los profesionales y la transparencia de los sistemas de IA (es decir, la descripción de su lógica, sesgo potencial y limitaciones) puede reducir las consecuencias negativas.

La transparencia también tiene que caracterizar el uso por parte de los tribunales de la IA para la investigación jurídica y el análisis de documentos. Los jueces han de ser transparentes en cuanto a qué decisiones dependen de la IA y cómo se utilizan los resultados proporcionados por la IA para contribuir a los argumentos, en consonancia con los principios de juicio justo e igualdad de armas.

Por último, la transparencia puede desempeñar un papel importante en relación con los chatbots jurídicos basados en IA, haciendo que los usuarios conozcan la lógica y los recursos utilizados (por ejemplo, la lista de casos analizados). La plena transparencia también debe incluir las fuentes utilizadas para entrenar estos algoritmos y el acceso a la base de datos utilizada para proporcionar respuestas.

Cuando estas bases de datos son privadas, es necesario que se disponga de auditorías de terceros para evaluar la calidad de los conjuntos de datos y cómo se han abordado los posibles sesgos, incluido el riesgo de infrarrepresentación o sobrerrepresentación de determinadas categorías (no discriminación).

Otras cuestiones críticas afectan a las aplicaciones de IA diseñadas para automatizar la resolución alternativa de litigios o para apoyar la decisión judicial. Aquí, la distinción entre justicia codificada y justicia equitativa sugiere que el uso de la IA debería limitarse, a efectos de toma de decisiones, a los casos caracterizados por evaluaciones rutinarias y basadas en hechos. Esto implica la importancia de llevar a cabo más investigaciones sobre la clasificación de los diferentes tipos de procesos de toma de decisiones para identificar aquellas aplicaciones rutinarias del razonamiento jurídico que pueden demandarse a la IA, preservando en cualquier caso la perspectiva humana que también garantiza la creatividad jurídica de los responsables de la toma de decisiones.

En cuanto a la justicia equitativa, su lógica es más complicada que el simple resultado de casos individuales. Valores y consideraciones expresados y no expresados, tanto jurídicos como no jurídicos, caracterizan el razonamiento de los tribunales y no son replicables por la lógica de la IA. Los sistemas basados en ML no son capaces de realizar un razonamiento jurídico, sólo extraen inferencias identificando patrones en conjuntos de datos jurídicos, que no es lo mismo que la elaboración de un razonamiento jurídico.

Teniendo en cuenta el contexto más amplio del papel social de los tribunales, la jurisprudencia es un sistema en evolución, abierto a nuevas cuestiones sociales y políticas. Por lo tanto, las herramientas de IA dependientes de la trayectoria anterior podrían obstaculizar este proceso evolutivo: la naturaleza deductiva y dependiente de patrones de ciertas soluciones de IA puede socavar el importante papel de los responsables humanos en la evolución del derecho en la práctica y el razonamiento jurídico.

Además, en el plano individual, la dependencia de la trayectoria en relación con los datos anteriores también puede entrañar el riesgo de ‘análisis deterministas’, lo que provocaría el resurgimiento de doctrinas deterministas en detrimento de doctrinas de individualización de la sanción y en perjuicio del principio de rehabilitación e individualización en la imposición de penas.

Además, en varios casos, incluidos los ADR, tanto la mediación entre las demandas de las partes como el análisis del componente psicológico de las acciones humanas (culpa, intencionalidad) requieren una inteligencia emocional que los sistemas de IA no poseen. En cuanto al problema de los sesgos, como afirma la Comisión Europea para la Eficacia de la Justicia, "la neutralidad de los algoritmos es un mito, ya que sus creadores les transfieren, consciente o involuntariamente, sus propios sistemas de valores". Numerosos casos de sesgo en relación con las aplicaciones de la IA confirman que estos sistemas ofrecen con demasiada frecuencia -aunque en muchos casos de forma no intencionada- una representación parcial de la sociedad y de los casos individuales, lo que no es compatible con los principios de igualdad de trato ante la ley y de no discriminación.

La calidad de los datos y otras formas de evaluación de la calidad (evaluación de impacto, auditorías, etc.) pueden reducir este riesgo pero, debido a la naturaleza de los intereses potencialmente afectados en caso de decisiones sesgadas, los riesgos siguen siendo elevados en el caso de la justicia equitativa y parecen desproporcionados con respecto a los beneficios, en gran medida en términos de eficiencia para el sistema judicial.

Otras preocupaciones están relacionadas con el derecho a un juicio justo y la igualdad de armas procesales, cuando las decisiones judiciales se basan en los resultados de algoritmos patentados cuyos datos de entrenamiento y estructura no están a disposición del público. Una noción amplia de transparencia podría abordar estas cuestiones en relación con el uso de la IA en las decisiones judiciales, pero la transparencia de la IA -un objetivo difícil en sí mismo- no puede abordar las demás objeciones estructurales y funcionales citadas anteriormente.

Además, los científicos de datos pueden dar forma a las herramientas de IA de diferentes maneras en las fases de diseño y formación, por lo que si las herramientas de IA se convirtieran en una parte obligatoria del proceso de toma de decisiones, los gobiernos que seleccionen las herramientas que utilizarán los tribunales podrían interferir indirectamente en la independencia de los jueces.[18] Por esta razón, tendrá un papel

[18] Este riesgo no queda eliminado por el hecho de que el juez siga siendo libre de no tener en cuenta las decisiones de la IA, lo que supone una motivación específica. Aunque la supervisión humana es un elemento importante, su impacto efectivo puede verse socavado por la propensión psicológica o utilitaria (rentabilidad) del responsable humano de la toma de decisiones a aprovecharse de la solución aportada por la IA.

central la componente relacionada con la protección de los derechos fundamentales en el contexto de la evaluación de conformidad exigida por el AI Act.

4. Conclusión

La última oleada de desarrollo de la IA está teniendo un impacto transformador cada vez mayor en la sociedad y plantea nuevas cuestiones en varios campos, desde la medicina predictiva y la moderación de los contenidos de los medios de comunicación hasta los sistemas judiciales.

El sector público tiene mucho interés en aprovechar de las nuevas herramientas de la IA bajo la presión debida a la falta de recursos adecuados in varios sectores y la necesidad de proporcionar servicios más eficaces.

Por otro lado, hay una posibilidad limitada de las entidades públicas de desarrollar de manera autónoma las herramientas de IA necesarias a diferentes niveles. A nivel alto de sistemas nacionales para la administración pública, de momento, las herramientas más poderosas – incluyendo los modelos generativos – son proporcionadas para un número limitado de grandes empresas, muchas veces non europeas. A nivel de implementación concreta y local, donde pueden jugar un papel también proveedores de menor tamaño, hay por otro lato una frecuente falta de capacidad técnica de las entidades públicas.

En este contexto, además de las consideraciones específicas formuladas en las secciones anteriores, se destaca claramente el papel de la nueva normativa europea sobre la inteligencia artificial. El AI Act, con su enfoque en los riesgos y con particular referencia al impacto sobre los derechos fundamentales, puede contrarrestar de manera eficaz las consecuencias negativas del uso del AI poniendo a cargo de los proveedores específicas obligaciones y también pidiendo a las entidades que utilizarán éstas herramientas de evaluar ulteriores riesgos específicos relacionados con el contexto de uso.

Todavía, porque el enfoque sobre el riesgo se convierta en una solución eficaz de protección y aún más de respecto de los derechos fundamentales por diseño, es necesario que las obligaciones de evaluación del riesgo se concreten en modelos viables de análisis y medición (en términos de escala de riesgo)[19] y que también se formen, tanto en el sector público como en lo privado, expertos capaces de evaluar y gestionar estos riesgos. Esto se debe a que la evaluación de impacto y su papel en situar el respecto de los

[19] Véase Mantelero (2022: 45-91).

derechos fundamentales en el centro del diseño de las herramientas de IA no pueden reducirse a ejercicios formales basados en formularios.

Además y finalmente, modelos adecuados de evaluación de riesgos con clara definición de los parámetros relevantes y de la manera de combinarlos en la evaluación de riesgos[20] proporcionan garantías también para los proveedores de IA y los usuarios principales en términos de rendimiento de cuentas y de capacidad de comprobar la natura adecuada de las soluciones adoptadas con respecto a la futura actividad de inspección y sanción que las autoridades competentes van desarrollando en el marco del AI Act.

Bibliografía

Agre, Philip E. (1994): "Surveillance and Capture: Two Models of Privacy", en The Information Society, núm. 10, pp. 101-127.

Brauneis, Robert y Goodman, Ellen P. (2018): "Algorithmic Transparency for the Smart City", en Yale J.L. & Tech., núm. 20, pp. 103-131.

Burrell, Jenna (2016): "How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms", en Big Data & Society, núm. 3, doi: 10.1177/2053951715622512.

Caruana, Rich et al. (2015): "Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission", en Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, Association for Computing Machinery pp. 1721–1730.

Crawford, Kate y Joler, Vladan (2018): "Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources", http://www.anatomyof.ai.

European Union Agency for Fundamental Rights (2019): Data Quality and Artificial Intelligence – Mitigating Bias and Error to Protect Fundamental Rights.

Eykholt, Kevin et al. (2018): "Robust Physical-World Attacks on Deep Learning Visual Classification", en *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), Salt Lake City, pp. 1625-1634, doi: 10.1109/CVPR.2018.00175.

Fjeld, Jessica et al. (2020): Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center for Internet & Society, Cambridge, MA, https://papers.ssrn.com/abstract=3518482.

Graber, Christoph B. (2020): "Artificial Intelligence, Affordances and Fundamental Rights", en M. Hildebrandt y K. O'Hara (ed.), Life and the Law in the Era of Data-Driven Agency, Cheltenham, Edward Elgar, pp. 194–213.

Hagendorff, Thilo (2020): "The Ethics of AI Ethics: An Evaluation of Guidelines", en Minds and Machines núm. 30, pp. 99-120.

[20] Véase la nota 19.

Hildebrandt, Mireille (2019): “Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning”, en Theoretical Inquiries in Law, núm. 20 (1), pp. 83-121.

Jobin, Anna et al. (2019): “The Global Landscape of AI Ethics Guidelines”, en Nature Machine Intelligence, núm 1, 389–399.

Loideain, Nóra N. y Adams, Rachel (2020): “From Alexa to Siri and the GDPR: The Gendering of Virtual Personal Assistants and the Role of Data Protection Impact Assessments”, en Computer Law & Security Review, núm. 36, doi:10.1016/j.clsr.2019.105366.

Mantelero, Alessandro (2016): “Personal Data for Decisional Purposes in the Age of Analytics: From an Individual to a Collective Dimension of Data Protection”, en Computer Law & Security Review, núm. 32(2)., pp. 238-255.

Mantelero, Alessandro (2022): Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI, The Hague, T.M.C. Asser Press-Springer.

Nickerson, Raymond S. (1998): “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises”, en Review of General Psychology, núm. 2, pp. 175-220.

Pasquale, Frank (2015): The Black Box Society. The Secret Algorithms That Control Money and Information, Cambridge, MA, Harvard University Press.

Raso, Filippo et al (2018): Artificial Intelligence & Human Rights Opportunities & Risks. Berkman Klein Center for Internet & Society, Cambridge, MA, https://cyber.harvard.edu/sites/default/files/2018-09/2018-09_AIHumanRightsSmall.pdf?subscribe=Download+the+Report.

Selbst, Andrew D. (2017): “Disparate Impact in Big Data Policing” en Georgia Law Review, núm. 52 (1), pp. 109-163.

Tubaro, Paola et al. (2020): “The Trainer, the Verifier, the Imitator: Three Ways in Which Human Platform Workers Support Artificial Intelligence”, en Big Data & Society, núm. 7, doi: 10.1177/2053951720919776

Veale, Michael y Binns, Reuben (2017): “Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data”, en Big Data & Society, núm. 4 (2), doi:10.1177/2053951717743530.

West, Sarah M. et al. (2019): “Discriminating Systems. Gender, Race, and Power in AI”, https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2.