SOBRE INTELIGENCIA ARTIFICIAL, DECISIONES JUDICIALES Y VACÍOS DE ARGUMENTACIÓN*

ON ARTIFICIAL INTELLIGENCE, JUDICIAL DECISIONS AND ARGUMENTATION VOIDS

Dyango Bonsignore Fouquet
Profesor Ayudante de Derecho penal
Universidad de Alicante

RESUMEN

En las últimas décadas, hemos asistido a una auténtica explosión del campo de la inteligencia artificial (IA), cuya popularización se ha hecho especialmente visible en tiempos recientes. Una de las aplicaciones de esta tecnología que más interés viene suscitando es su empleo en el ámbito jurídico. Este artículo pretende reflexionar sobre una parte reducida de la conexión Derecho-IA, a saber, aquella que concierne a la toma de decisiones judiciales. Se examinan, para ello, dos hipótesis: la que concibe la sustitución del juzgador humano por una inteligencia artificial, y la que entiende el uso de esta última como complemento o apoyo al juzgador a lo largo del proceso decisorio. En relación con la primera, se reflexiona sobre las dificultades que entraña el concepto de «razón» cuando se aplica al decisor humano y al artificial, y en qué modo esto repercute en la función jurisdiccional como actividad *inter pares*. En relación con la segunda cuestión, se destaca el potencial que el uso de la IA como apoyo puede venir acompañada de sus propias dificultades. Reflexionando a través del ejemplo de la justicia penal y los pronósticos de reincidencia, se destaca que el uso de este tipo de instrumentos puede producir «vacíos de argumentación», a saber, puntos ciegos en el proceso de justificación de la decisión judicial.

PALABRAS CLAVE

Inteligencia artificial, algoritmos, *machine learning*, opacidad decisión judicial, argumentación jurídica.

ABSTRACT

In recent decades, we have witnessed a veritable explosion in the field of artificial intelligence (AI), the popularization of which has become particularly visible in recent times. One of the most interesting applications of this technology can be found in the legal field. This article aims to reflect on a small part of the Law-IA relationship, namely that which concerns judicial decision-making. To this end, two hypotheses are examined: one that envisages the replacement of the human judge by an artificial intelligence, and another that understands the use of the latter as a complement or support for the judge throughout the decision-making process. Regarding the first issue, it reflects on the difficulties posed by the concept of «reason» when applied to the human and artificial decision-maker, and how this impacts on the jurisdictional function as a peer-to-peer activity. In relation to the second question, the paper highlights the potential that the use of AI as a support can come with its own difficulties. Reflecting through the example of criminal justice and recidivism predictions, it is argued that the use of such tools can produce «argumentation voids», i.e., blind spots in the justification of the judicial decision.

KEY WORDS

Artificial Intelligence, algorithms, machine learning, opacity, judicial decision-making, legal argumentation.

DOI: doi.org/10.36151/td.2021.011

* Este trabajo se inscribe en el marco del proyecto «Manifestaciones de desigualdad en el actual sistema de justicia penal: examen crítico de las razones de necesidad, oportunidad y peligrosidad para la diferencia (AEQUALITAS)» (RTI2018-096398-B-I00) del Ministerio de Ciencia e Innovación. El autor desea expresar su agradecimiento a todos aquellos que han contribuido a la elaboración de este texto con sus observaciones a versiones previas del mismo, agradecimiento extensible a los revisores anónimos cuyos comentarios han contribuido a la mejora sustancial del texto..

SOBRE INTELIGENCIA ARTIFICIAL, DECISIONES JUDICIALES Y VACÍOS DE ARGUMENTACIÓN

Dyango Bonsignore Fouquet
Profesor Ayudante de Derecho penal
Universidad de Alicante

1. INTRODUCCIÓN

A lo largo de las últimas décadas se ha difundido la idea de que asistimos a una revolución científica y social sin precedentes. El desarrollo en materia de tecnologías de la información y la comunicación en general, y en ciencias de la computación en particular, ha sido fulgurante. Tanto es así que la transición de la «ciencia ficción» a la ciencia *tout court* casi parecería haberse producido sin solución de continuidad. Si no hace tanto apenas era posible imaginar con cierto grado de detalle la posibilidad de *inteligencias artificiales*, en la actualidad ya es posible encontrarlas en espacios tan banalizados como la electrónica de consumo.

Hablar de inteligencia artificial (IA), sin embargo, resulta manifiestamente impreciso en el marco del vocabulario sociotecnológico contemporáneo, tan variado como mal delimitado. Así, frecuentemente hablamos de IA en conjunto o solapadamente con otras nociones como *Big Data*, *machine learning*, algoritmos o similares (Miró Llinares, 2018: 91). El carácter intercambiable que en ocasiones presentan estos términos en el uso común

es revelador, pues da cuenta de una serie de transformaciones tecnológicas que han cobrado su actual relevancia social precisamente porque se han presentado unidas. En plena consonancia con los cánones contemporáneos, la expresión «inteligencia artificial» ya no designa la misma realidad que en la década de los 50 (Grosan y Abraham, 2011: 3). Con todo, ese sentido laxo e indiferenciado resulta, por el momento, suficiente para los propósitos de esta discusión inicial.

En suma, se diría que actualmente toda actividad de cierta complejidad puede tener algo que ganar con la introducción de este tipo de instrumentos computacionales, desde la gestión del tráfico a la participación en los mercados financieros, pasando por la propia investigación científica (Russell y Norvig, [3]2010: 28-9). En este trabajo, sin embargo, interesa tan solo uno de los ámbitos en que se ha explorado la aplicabilidad de la IA, a saber, su empleo en el ámbito del Derecho. Aun dentro de este ámbito, el análisis se ceñirá a una parcela relativamente acotada: el ámbito de la toma de decisiones jurídicas (y, singularmente, de las decisiones judiciales).

El planteamiento es el siguiente: los avances en inteligencia artificial han permitido situar potentes algoritmos en el corazón de múltiples procesos complejos de toma de decisiones. Podría decirse, incluso, que un objetivo central de la investigación en la materia se orienta a la construcción de «decisores» artificiales que se valdrían de la capacidad de procesar grandes cantidades de información para llevar a cabo dicha tarea de manera más eficaz y eficiente que sus homólogos humanos. Sin embargo, esto resulta especialmente problemático cuando se trata de extender al campo de la toma de decisiones que realiza, por ejemplo, un juez. Anticipando un tanto la discusión, cabe señalar que impartir justicia es algo más que *decir* el Derecho, y que en la labor de *dictarlo* se requiere un actor humano.

La hipótesis del «juez IA» (Sourdin, 2018: 1117), es decir, la sustitución del juez por un algoritmo, no es sin embargo la única manifestación posible de la automatización de la toma de decisiones jurídicas. También cabe pensar que herramientas computacionales como las que nos ocupan pueden servir como instrumento del proceso de toma de decisión llevado a cabo por el juez (u otro operador jurídico). En este caso, sin embargo, emergen otras dificultades, entre las que se encuentra el polémico problema de la «opacidad» del algoritmo (Burrell, 2016). A tal efecto, parece que entender las decisiones judiciales como un producto de procesos argumentativos *justificativos* se compadece mal con la posibilidad de apoyarse en instrumentos cuyo funcionamiento sea inaccesible al escrutinio externo. La posibilidad de que el uso de estos dispositivos genere *vacíos argumentativos* merece, por tanto, ser examinada algo más atentamente.

Para abordar esta última problemática, se propone tomar como asidero los protocolos para la valoración del riesgo delictivo empleados en sede judicial. Lo que aquí interesa no es llevar a cabo una reflexión procesal sobre el papel de este tipo de instrumentos ni un examen minucioso de los aspectos técnicos de los mismos, sino más bien enfocar la discusión desde la función argumentativa de la decisión judicial. Determinar si tales protocolos son piezas satisfactorias para construir la justificación de una decisión, qué requisitos han de cumplir para ello o, incluso, si es posible leer las valoraciones protocolizadas del riesgo como argumentos en sí mismos será el objeto de discusión en lo sucesivo. En este sentido,

el trabajo adopta una óptica de análisis que atribuye un papel central a la justicia penal, a la que recurre como plataforma para discutir cuestiones de fondo que, tal vez, pudieran resultar transversales.

En definitiva, se pretende prestar atención a los problemas que emergen cuando, sobre el mismo objeto jurídico, confluyen dos «decisores» diferentes. Por un lado, el juez racional, que adopta decisiones basadas en razones justificadas argumentativamente. Por otro, la inteligencia artificial, que toma decisiones sobre la base de patrones, correlaciones y heurísticos no siempre transparentes y de implicaciones jurídicas aún inciertas.

2. INTELIGENCIA ARTIFICIAL. BREVES NOCIONES CONCEPTUALES

Como se ha avanzado, hablar de inteligencia artificial es, hoy por hoy, poco más que una constatación de hechos consumados. De hecho, y aunque las primeras manifestaciones de este tipo de tecnologías se remonten a mediados del siglo pasado (Russell y Norvig, [3]2010: 16-18), parece que solo recientemente se ha alcanzado la suficiente sofisticación técnica como para transformar la IA en un acontecimiento sociotecnológico transversal. Ejemplos pintorescos de un fenómeno cuya magnitud resulta difícil de estimar son, por ejemplo, los concursos de belleza con un «jurado IA»[1] (Levin, 2016) o la venta de un cuadro elaborado por un algoritmo «artista» (Cohn, 2018).

Sin embargo, la irrupción de la inteligencia artificial se ha producido de forma confusa (Edwards y Veale, 2018: 1), en la medida en que ha podido cobrar todo su significado dentro de un contexto de desarrollo sociotecnológico más amplio. La informatización de la vida cotidiana se ha acelerado a través de la multitud de dispositivos cada vez más *personales* que acompañan a los individuos y los insertan en una red en constante expansión. El carácter utilitario que tuvieron la informática y la electrónica ha dado paso, hace ya un tiempo, a *un ethos hedonístico* que ha transformado la interacción persona-dispositivo en una fuente de disfrute. La premisa subyacente al desarrollo tecnológico contemporáneo parece apuntar a la convergencia entre lo analógico y lo digital, algo que expresan bien algunos términos del incipiente «vocabulario ciborg» como *realidad aumentada*.

En este contexto de relativa confusión, resulta indispensable clarificar, pues, qué es lo que necesitamos saber.

2.1. INTELIGENCIA ARTIFICIAL Y ALGORITMOS

Desde un punto de vista intuitivo/coloquial, la expresión inteligencia artificial hace referencia, de forma más bien figurada, al modo en que determinadas aplicaciones informáticas procesan la información de un modo automatizado, produciendo resultados que denotan inteligencia (Surden, 2014: 90). Sin embargo, desde un punto de vista algo más técnico, con el término inteligencia artificial nos referimos a un campo de investigación y

desarrollo en el que convergen diversas ramas de conocimiento y cuyo propósito «[…] no es solo comprender, sino construir entidades inteligentes» (Russell y Norvig, [3]2010: 1).

Aunque aquí interesa específicamente la IA como instrumento (y no como campo), no hay una única forma de replicar inteligencia de manera artificial[2]. Ahondando en la confusión, dos términos que se han empleado usualmente como intercambiables son *IA* y *algoritmo*. En realidad, la relación entre uno y otro sería más bien del todo a la parte: el comportamiento de una IA puede explicarse en buena medida a través de su correspondiente algoritmo, dado que este expresa el modo en que la información es procesada hasta producir una determinada respuesta. Sin embargo, hablar de algoritmos no comporta necesariamente hablar de inteligencia artificial, dado que, en su sentido nuclear, los algoritmos son simplemente reglas sistemáticas para la solución de problema que, sobre la base de determinadas premisas, conducen a una conclusión (Berk, 2018: 32). La IA implica, por tanto, el uso de algoritmos capaces de interactuar con su entorno (estrictamente, con los datos generados por dicho entorno), *aprender* y *adaptarse.*

Un ejemplo relativamente trivial de este proceso es el de los filtros anti-*spam.* Este tipo de programas tienen en su base un algoritmo cuyo funcionamiento, en su versión más elemental, resulta bastante intuitivo. Para distinguir el correo legítimo y correo no deseado, se extraen una serie de palabras clave que aparecen frecuentemente en la segunda categoría de correos y, sobre esa base, se aplica el filtro a los mensajes futuros. Este sería, sin embargo, un algoritmo sencillo de esquivar, pues bastaría evitar la introducción de las palabras discriminantes. Sin embargo, la situación cambia cuando el usuario señala manualmente nuevos correos no deseados que se han «escapado» del filtro para que sean incorporados a la base de datos. Esto conduce a modificar el catálogo de palabras discriminantes, así como a modificar el peso que tienen las ya detectadas para clasificar un nuevo correo como *spam* (Burrell, 2016:7-9).

Esta segunda parte del procedimiento es la que sirve de ejemplo para el tipo de algoritmos que nos ocupa, es decir, aquellos que poseen la capacidad de modificar sus elementos automáticamente a través de la incorporación de nueva información. Esto nos conduce a la idea de *machine learning*, así como a considerar la importancia de los datos y su cantidad en la construcción contemporánea de inteligencia artificial.

2.2. *BIG DATA* Y *MACHINE LEARNING*

En efecto, como se apuntó en las líneas iniciales del texto, resulta difícil de comprender la actual discusión sobre la IA sin hacer referencia a la interacción que se produce entre *machine learning* y *Big Data.* Veámoslo.

En términos elementales, cuando hablamos de *machine learning* (al igual que cuando hablamos de IA) podemos referirnos bien a un campo de las ciencias de la computación, bien al producto de dicho campo, a saber, a aquellos programas capaces de aprender de la experiencia y mejorar su rendimiento sobre la base de dicho aprendizaje. La idea de aprendizaje se emplea, con todo, en un sentido fundamentalmente metafórico respecto del aprendizaje humano (Surden, 2014: 89). Desde un punto de vista «conductista», se en-

tiende que el algoritmo «aprende» en la medida en que modifica su comportamiento para adaptarse mejor a los datos. Esto hace que, conforme la base de datos aumenta, en principio el rendimiento del algoritmo mejora. Pero este aprendizaje no depende únicamente de la incorporación de datos, sino también de la capacidad de este tipo de programas para inferir patrones (correlaciones) entre los datos y extraer reglas de carácter general de acuerdo con un procedimiento estrictamente inductivo. Por todo ello, la mayoría de algoritmos de *machine learning* incorporan dos procesos paralelos: un clasificador que agrupa los datos brutos en categorías (sobre la base de palabras clave, clasifica como *spam*/no *spam*); y un *learner* encargado de «entrenar» al algoritmo a través de la incorporación de datos (Burrell, 2016: 5).

No obstante, la auténtica revolución en materia de inteligencia artificial tal vez no ha procedido tanto de la sofisticación de los algoritmos cuanto de su desarrollo al calor de un fenómeno paralelo que ha dado en llamarse *Big Data*. Con la expresión *Big Data* suele aludirse a la disponibilidad de una ingente cantidad de datos, en constante expansión, fruto del estado sociotecnológico de las sociedades contemporáneas (someramente descrito *supra*). Sin embargo, la referencia a la cantidad de datos no es seguramente suficiente. Aquello que completa la caracterización del *Big Data* es la *interoperabilidad* de los datos, es decir, la posibilidad de agregarlos, combinarlos y manipularlos. Por esta razón, se ha afirmado que aquello que recibe el nombre de *Big Data* se entiende mejor como un fenómeno cultural, tecnológico y «académico» basado en la interacción de *i)* una *tecnología* capaz de generar y manipular ingentes cantidades de información; *ii)* la importancia del *análisis* de dicha información para la toma de decisiones sobre asuntos socialmente relevantes; y *iii)* una *mitología* que confiere a estas nuevas formas de aproximación al conocimiento un aura de objetividad previamente inalcanzable (Boyd y Crawford, 2012: 663).

La relación simbiótica entre los desarrollos en materia de IA y este estado de cosas sociotecnológico resulta evidente (Edwards y Veale, 2017: 25). Tanto es así que entre los expertos ha calado la percepción de que lo determinante para obtener algoritmos eficaces (o lo que es lo mismo, para aproximarse a la obtención de una *auténtica IA*) no es lidiar con el programa, sino insuflarlo con suficientes datos y dejar que el aprendizaje autónomo haga el resto (Russell y Norvig, [3]2010: 27-8). Así las cosas, tan solo parece necesario añadir potencia computacional y materia prima (datos) para transformar *más* en *mejor*.

3. LAS RELACIONES ENTRE EL DERECHO Y LA INTELIGENCIA ARTIFICIAL

Los ámbitos del Derecho y el desarrollo en inteligencia artificial han encontrado tradicionalmente amplias zonas de interés común. El carácter normativamente pautado del proceder jurídico, así como su orientación práctica hacia la toma de decisiones y la resolución de conflictos parecía ser una esfera privilegiada para poner a prueba la IA. Las aplicaciones, en este sentido, han sido múltiples: desde la modelización de argumentos jurídicos bajo forma computacional (*v. gr.* Li *et al.*, 2018), la predicción de los resultados de sentencias (*v. gr.* Aletras *et al.*, 2016; Medvedeva, Vols y Wieling, 2019) o, simplemente, la agilización de

la actuación de las partes en el procedimiento (para una panorámica sobre la investigación en IA y Derecho, sin ánimo de exhaustividad, *vid*. Barona Vilar, 2021: 555-63; Belloso Martín, 2019: 2-3; Bench-Capon y Dunne, 2007; Feteris, 2017: 33-41; Medvedeva *et al.*, 2019: 1-6; Nieva Fenoll, 2018; Rissland, Ashley y Loui, 2003).

En esta materia, la argumentación jurídica ha sido un área de interés frecuente y tal vez recíproco. Por un lado, la tradicional discusión sobre las posibilidades de formalización de los argumentos jurídicos a través de expresiones lógicas encontraba en la informática una disciplina suficientemente sofisticada y tecnificada como para introducir un elemento experimental con el que avanzar en el debate. Por otro, los desarrolladores de algoritmos «inteligentes» encontraban en el razonamiento judicial el paradigma de la toma de decisiones complejas basadas en razones. En una singular confluencia, por tanto, si desde la lógica jurídica se buscaba imprimir al razonamiento jurídico objetividad matemática, la investigación en IA buscaba dotar de algo más de humanidad al algoritmo.

3.1. LA DIMENSIÓN ARGUMENTATIVA DEL DERECHO Y LA IA

Sentado el contexto y reflejada —someramente— la prolífica relación entre inteligencia artificial y Derecho, el presente texto no podría dar cuenta de la totalidad del trabajo realizado en este campo. Por tanto, procede llevar a cabo una primera restricción de la temática a abordar. La atención se dirigirá a la interacción entre IA y argumentación jurídica en el marco de la decisión judicial. Incluso dentro de este ámbito, no se examinarán, por ejemplo, las distintas formas de modelado computacional de las decisiones judiciales ni los esfuerzos por realizar predicciones del comportamiento de los jueces sobre la base de dichos modelos. Dicho de otro modo, aquí no interesa la argumentación como «técnica» de producción de argumentos. Tampoco los argumentos-producto, al menos en lo que concierne a su estructura. En cambio, sí interesa la dimensión «sustancial» de la labor argumentativa, a saber, su papel integral en el funcionamiento y la efectiva materialización del Derecho, es decir, la argumentación jurídica como *actividad institucional*, podría decirse a riesgo de que la expresión resulte excesivamente general.

En este sentido, la argumentación se encuentra en el centro de la interacción entre el legislador y la ciudadanía, a través del juzgador y las partes del proceso. El punto de interrogación esencial es la indagación sobre el modo en que la irrupción de la IA puede afectar a la toma de decisiones judiciales y, más en profundidad, repercutir sobre la legitimidad de un Derecho asentado en una forma de «hacer justicia» fundamentalmente argumentativa (Atienza, 2013: 28-31).

La relación entre la argumentación jurídica y la inteligencia artificial, si no se presupone, sí se ve especialmente favorecida por un determinado estado de cosas jurídico. En este sentido, la creciente importancia del proceder argumentativo en el Derecho, frente a la clásica representación del juez *bouche de la loi* confiere a las decisiones judiciales el tipo de complejidad que interesa en el desarrollo de algoritmos inteligentes. La cosmovisión del formalismo jurídico, cabe pensar, habría sido mucho menos fructífera a la hora de producir el tipo de «sinergia» disciplinaria de que hablan Rissland *et al.* (2003). Así, y si bien es

cierto que razonamiento jurídico e IA encuentran en la lógica clásica una raíz común, no ha sido sino a partir del llamado «giro argumentativo»[3] en el propio campo «Derecho-IA» cuando se ha abierto la veda paraa un entendimiento moderno de la cuestión (Bench-Capon y Dunne, 2007: 633-4).

3.2. DECISIÓN JUDICIAL, ARGUMENTACIÓN E INTERPRETACIÓN

Es bien conocido que decidir y argumentar no son términos equivalentes. Ni toda decisión deriva de un procedimiento argumentativo previo ni toda argumentación procede necesariamente de una situación que requiera decidir en un sentido estricto o práctico del término (Atienza, 2013:108). Con todo, lo expuesto en las líneas precedentes permite dejar de lado, siquiera temporalmente, esta problemática para el caso que nos ocupa. Y ello en la medida en que, circunscribiéndonos al tipo específico de decisiones adoptadas por el juez en el ejercicio de sus funciones, podemos asumir que nos hallamos ante decisiones argumentadas. El argumento es lo que permite al juzgador *justificar* su toma de posición institucional a través de una defensa razonada de los pasos que conducen a la misma (García Amado, 2016: 51).

En este sentido, las decisiones judiciales pueden ser notablemente complejas, incluso en los llamados «casos fáciles» (Taruffo, 1998: 311), pues requieren tomar en consideración multitud de requisitos normativos (procedimentales o sustantivos), de racionalidad en sentido amplio (*v. gr.* sobre la valoración de las circunstancias del caso, sobre relaciones lógicas…), o axiológicos, todos ellos, además, frecuentemente entrelazados.

Al final de este proceso, la legitimidad de la decisión procederá de su capacidad para cumplir con las exigencias de completud, consistencia, adecuación y coherencia de la justificación. En suma, deberá estar basada en «buenas razones» (Taruffo, 1998: 314-315). En este punto se intuye ya uno de los principales puntos de fricción entre la toma de decisiones «tradicional» y la «algorítmica». Si la IA hubiera de sustituir al decisor humano (al menos en el ámbito en que nos encontramos), parece que deberían cumplirse algunos requisitos (no exhaustivos): no solo habría de ser capaz de explicitar las razones que han conducido a la decisión, sino que estas razones deberían, además, ser significativas para las personas (Edwards y Veale, 2017: 53) —cuestión que retomaremos más adelante— y, por último, tener la legitimidad suficiente como para persuadir/convencer a los destinatarios o, a la inversa, ser suficientemente convincentes como para resultar legítimas (Atienza, 2017: 45; García Amado, 2016: 52). En este sentido, la capacidad de la decisión «algorítimica» para resultar convincente y legítima probablemente no dependa únicamente del contenido o la forma del argumento que enuncie, sino también del carácter autoritativo conferido a la IA por la sociedad.

Dicho lo anterior, siempre existe una posibilidad alternativa. Si la inteligencia artificial se considera inadecuada para tomar de manera autónoma determinadas decisiones (por ejemplo, dictar sentencia), siempre cabe la posibilidad de que intervenga como parte/complemento de la decisión del juez tradicional. En tal caso, la cuestión a dilucidar es la siguiente: cómo puede reconducirse el tipo de proceso decisorio de un algoritmo a las exi-

gencias de justificación que se requieren del juez. Nuevamente, la posibilidad de convertir los cálculos en razones sustanciales resulta indispensable, si bien aparece de modo más sutil que en el caso anterior. La posibilidad de que un procedimiento automatizado de este tipo intervenga como un punto más dentro de las circunstancias valoradas por el juez corre el riesgo de diluir las demandas de transparencia y justificación, aun cuando el resto de aspectos del argumentario judicial resulte razonable.

En las siguientes dos secciones se abordarán con mayor grado de detalle algunas de las implicaciones subyacentes a ambos usos de algoritmos inteligentes: la suplantación (del) y el apoyo al juez tradicional[4].

4. LA HIPÓTESIS DEL «JUEZ IA»

La primera hipótesis que cabe explorar es la del «juez IA» (Sourdin, 2018: 1115), es decir, la posibilidad de que la IA reemplace al juez en la toma de decisiones. Cabe introducir una primera matización al respecto: aunque el grueso de la reflexión que aquí se propone tenga que ver, explícitamente o no, con la «sentencia» como paradigma del ámbito decisorio de los jueces, no es infrecuente que la actividad judicial implique la toma de decisiones de menor alcance (instrumentales o no a aquella). La complejidad y el alcance de ambas cuestiones serán frecuentemente distintos, y es posible imaginar que algunos segmentos eminentemente rutinarios/burocráticos de la profesión pudieran ser automatizados (Sourdin, 2018: 1118). En este caso, «programar» y «decidir» podrían ser procesos sustancialmente idénticos (Martínez García, 2019: 171-2). No obstante, es probable que este tipo de situaciones no sean tan frecuentes ni, tal vez, tan ordinarias como pudiera pensarse (Taruffo, 1998:318).

La viabilidad del proyecto del «juez IA», por tanto, no puede depender de su aplicación a esta clase de casos. Tampoco, seguramente, del grado de acierto en las predicciones efectuadas sobre las resoluciones de jueces y tribunales (Aletras *et al.*, 2016; y Li *et al.*, 2018). Poder *predecir* no quiere decir poder *replicar*, y elaborar pronósticos acertados sobre las decisiones judiciales no nos informa más que vagamente del procedimiento seguido para llegar a ellas. Debe recordarse, además, que la posibilidad de realizar predicciones en Derecho resulta menos sorprendente de lo que puede parecer, en la medida en que hablamos de un campo significativamente reglado (Nieva Fenoll, 2018: 58-60). La consistencia en la resolución de un caso a otro similar y la formación de expectativas razonables al respecto son elementos característicos de un Derecho mínimamente funcional. Complementariamente, cierto margen de incertidumbre es igualmente consustancial a un Derecho entendido como *praxis* (Atienza, 2013: 108; García Amado, 2016: 50), moderadamente dinámico y en el que la solución no depende solo de la subsunción mecánica del caso en la regla.

Así las cosas, lo que se requiere de una IA judicial es que funcione adecuadamente en *casos difíciles* y, además, que lo haga de manera *razonada* o, al menos, reconducible a razones aceptables. La primera cuestión ha sido vista tradicionalmente con cierto escep-

ticismo (Taruffo, 1998: 318-9; y Sourdin, 2018: 1123). El proceso que conduce desde una controversia jurídica hasta su resolución pasa por múltiples dificultades y decisiones parciales, que involucran, a su vez, cuestiones de hecho y de Derecho vinculadas a reglas y principios y sobre las que normalmente caben, al menos dos interpretaciones razonables (Taruffo, 1998: 312-3). Ello implica, a su vez, una interacción constante de dos fuerzas en tensión: la primera, de carácter «subsuntivo» (en sentido laxo), implica elementos de juicio preestablecidos (una suerte de *tópica* conformada por la norma, criterios jurisprudenciales, máximas de experiencia…) que conectan el caso concreto con pautas de respuesta; y la segunda, de carácter «distintivo», tiende a extraer de las circunstancias particulares del caso motivos para exceptuar o modular esa respuesta preestablecida y considerada inadecuada. Este segundo aspecto del razonamiento judicial parece especialmente problemático cuando hablamos de inteligencia artificial, en la medida en que se encuentra estrechamente relacionado con valores, así como con una cierta permeabilidad del juez a los cambios de la sociedad que le rodea y de la que forma parte (Atienza, 2013: 29). La forma en que suelen construirse los algoritmos inteligentes que tratan de emular la labor judicial (fundamentalmente, a través de la incorporación del marco normativo y el *case law* disponible) le otorga un carácter marcadamente retrospectivo que genera dudas sobre su capacidad para responder adecuadamente a las singularidades de un caso novedoso (Barona Vilar, 2021: 555-6; Edwards y Veale, 2018: 1; Nieva Fenoll, 2018: 99; y Sourdin, 2018:1125).

Llegados a este punto, procede plantear si son, propiamente, *razones* aquello que podemos esperar de la decisión tomada por una IA. Para ello, resulta necesario detenerse nuevamente en la noción metafórica de inteligencia que se emplea en estos casos. Conviene abordar cómo *piensa* un algoritmo.

4.1. ¿CÓMO PIENSAN LAS MÁQUINAS? ALGORITMOS, Y HUMANOS, *INTELIGENTES*

Es habitual reconocer que las diferencias entre el pensamiento humano y el «computacional» no son solo de orden cuantitativo (en términos de rapidez de procesamiento, cantidad de datos, etc.) sino también, y principalmente, de orden cualitativo. A esto se refería la afirmación de Searle de que las máquinas poseen sintaxis, no semántica, en su célebre experimento mental de «La Habitación China» (Searle, 1980 y 2002). Asumir que la posición de Searle es correcta (omitiendo el complejo debate en torno a la cuestión) implica aceptar que, mientras sean digitales, las inteligencias artificiales no podrán llegar a ser *inteligentes* en un sentido profundo del término, pues están abocadas a carecer de estados mentales. Por muy sofisticada que pueda ser, la IA es resultado de un complejo entramado de relaciones sintácticas/formales, no semánticas/de contenido. Ello no quiere decir que no sean capaces de desarrollar procesos decisorios complejos, o incluso de mejorar su programación de manera autónoma, sino que, por mucho que puedan *simular* un comportamiento inteligente, no pueden llegar a *duplicar* la inteligencia humana (Searle, 2002: 273).

Esta forma «sintáctica» de pensar se observa ya claramente en tareas sencillas para las que se emplean técnicas de *machine learning* como el reconocimiento caligráfico. De acuer-

do con la explicación de Burrell (2016:5-7), el método de reconocimiento de patrones empleado por el algoritmo funciona similarmente al anteriormente descrito del *spam*: buscando repeticiones. Así, lo determinante es la frecuencia e intensidad con que algunas partes del trazo se repiten en las distintas iteraciones. En términos de procesamiento informatizado, esto implica *i)* digitalizar el símbolo manuscrito de forma que pueda representarse por píxeles y *ii)* otorgar un valor numérico a cada píxel según su capacidad para resultar discriminante. El conjunto de valores otorgados a cada píxel será mayor cuanto más relevante sea para detectar qué símbolos fueron escritos por la misma persona.

Lo que resulta significativo es que, en todo este procedimiento, no interviene ningún proceso característicamente humano de pensamiento. El programa no piensa por líneas, círculos, diagonales, etc., sino por puntos (píxeles, en este caso) que tienen asociados un valor numérico conforme a una regla de optimización. Se trata de un procedimiento de naturaleza estadística cuya interpretación humana puede ser considerablemente difícil y, frecuentemente, irreconducible a un plano de significados (Surden, 2014: 96-7; Edwards y Veale, 2017: 59). Retomando el ejemplo del *spam*, son las repeticiones de palabras o de grupos de palabras en los correos no deseados las que determinan su clasificación, y no el contenido del mensaje transmitido, la relación que este tiene con el remitente o el contexto interpretativo en el que se desenvuelve la comunicación. Esto permite ver de qué manera, incluso cuando es posible interpretar el procedimiento seguido por el programa para efectuar las clasificaciones, el proceso no tiene por qué conducir a resultados esclarecedores o que *signifiquen* algo relevante. Saber que una serie de palabras clave se repite con frecuencia en los correos no deseados no resulta demasiado satisfactorio cuando observamos el carácter mundano de muchas de ellas. El algoritmo clasificador nos dice que un correo es *spam* en virtud de las palabras que contiene y de su semejanza con las palabras contenidas en otros correos previamente clasificados, pero no nos dice *por qué* resulta indeseado. Encontrar una explicación es, en estos casos, una tarea *ad hoc* que requiere la imposición de un proceso interpretativo (conforme a cánones humanos de razonamiento) a un cálculo matemático de optimización estadística (Barona Vilar, 2021: 558-9; y Burrell, 2016: 8-9).

Visto lo anterior, procede volver sobre la idea de inteligencia que atribuimos a la IA. Desde luego, parece que este tipo de programas pueden producir *resultados* inteligentes e, incluso, mostrar *comportamientos* asociados a la inteligencia como aprender de la experiencia. No obstante, el *procedimiento* por el que el programa lleva a cabo dichas tareas no se basa en la replicación de los procesos cognitivos humanos, sino en el empleo de heurísticos (como las correlaciones estadísticas) que permitan llegar a resultados semejantes soslayando el elemento cognitivo/reflexivo. Una forma de «inteligencia por *proxy*», en términos de Surden (2014: 97).

Esto evidencia una concepción fuertemente funcionalista y utilitarista del funcionamiento de la IA en la toma de decisiones. Su finalidad no es replicar el pensamiento humano (resulta dudoso que pudiese hacerlo), sino ser igual de eficaz, o mejor, serlo en sus consecuencias. Y esto, precisamente, resulta problemático cuando tratamos de trasladar su aplicación a un contexto en que el intercambio de significados y la pugna por hacer prevalecer las definiciones/interpretaciones propias sobre el resto de puntos de vista juega un papel esencial.

4.2. LAS RAZONES DE LOS ALGORITMOS

Llegados a este punto, cabe plantearse si los algoritmos son capaces de producir el tipo de razones que se requieren para llevar a cabo la labor jurisdiccional. A este respecto, trasladar en términos computacionales aquello que es jurídicamente relevante para la resolución de controversias en áreas fuertemente reglamentadas puede, *a priori*, concebirse como viable. Y ello no solo porque sea más sencillo realizar la traducción norma-programa, sino porque generalmente las razones requeridas para la justificación no serán otras que el mandato del respeto a la ley. Constatar que el sujeto infringió la norma (por ejemplo, haber circulado a una velocidad excesiva) es todo el razonamiento que se requiere y que se está dispuesto a ofrecer. En la mayoría de los casos, además, resulta incluso preferible que este sea el funcionamiento por defecto, siquiera por cuestiones de operatividad. Aquí, un «juez IA» parece *prima facie* viable.

Sin embargo, este caso no es el paradigmático de la discusión IA-Derecho ni llama suficientemente la atención sobre el problema de las razones. ¿Puede una IA acompañar sus decisiones de razones *significativas* en casos que sí requieren entrar al fondo del asunto? En este sentido, nuevamente, cabe pensar en algoritmos capaces de aprender, a través de procedimientos como los descritos, el tipo de patrones de razonamiento que vienen asociados a la resolución de ciertos casos. Un algoritmo que pudiera recopilar y agrupar las «locuciones argumentales» más usuales para cada caso. Pero esto no sería diferente a la detección de palabras clave de un filtro anti-*spam*. Un catálogo de circunstancias C1, C2, C3… asociado a una respuesta R (digamos, absolver o condenar) estaría estadísticamente correlacionado, a su vez, con ciertos argumentos A1, A2, A3… Esto no sería más que extraer patrones de los múltiples razonamientos previamente efectuados por jueces y tribunales[5].

Admitamos, siquiera como ejercicio mental, que todo este procedimiento puede llevarse a buen puerto y que puede alcanzarse una *simulación* convincente del razonamiento judicial. Esta sería capaz de incorporar una conexión discursiva entre circunstancia fáctica, norma, y decisión a través del recurso a un argumentario, estadísticamente seleccionado, pero percibido como relevante desde el punto de vista de su significado. En otros términos, la IA debería ser capaz de satisfacer, no ya el test de Turing, sino al menos lo que podríamos llamar el «test de Toulmin». Es decir, debería ser capaz de producir argumentos que, por incorporar las mínimas características exigibles[6], resultaran indistinguibles de aquellos que construiría un juez humano.

Con todo, incluso en un caso semejante no parece que el resultado sea plenamente satisfactorio. Por un lado, es difícil de armonizar con la idea de ponderación de principios y el uso de los mismos en la argumentación jurídica. Como mucho, estos podrían tener un reflejo si el algoritmo fuera capaz de detectarlos en el *case law* e incorporarlos a su respuesta. Además, existe otro problema que tiene que ver no ya con el argumento y su contenido, sino con el tipo de función que cumple la argumentación en el seno de la práctica social de aplicar la norma al caso concreto. En la coloquial expresión «hacer justicia» anida un componente más profundo de interacción comunicativa entre las personas intervinientes en el proceso, así como entre las instituciones y la sociedad civil, sobre el que procede realizar una breve reflexión tentativa a modo de cierre.

4.3. EL ELEMENTO ANTROPOLÓGICO

Decíamos que las decisiones jurisdiccionales se toman en contextos comunicativos complejos en los que la interacción entre los intervinientes queda atravesada por los respectivos roles institucionalizados de cada uno de ellos. Conviene examinar esto con algo más de detenimiento.

Desde luego, cabe argüir que, en muchos casos, las controversias jurídicas pueden incorporar diversos intereses distintos al de hacer prevalecer una visión determinada de lo «justo». La actividad jurisdiccional puede ser instrumental, utilitaria y relativamente desapasionada. Sin embargo, resolver un conflicto será, con frecuencia, mucho más que un ejercicio rutinario de adjudicación de la razón a la parte mejor sustentada por el marco normativo preexistente. Los conflictos en sentido fuerte (aquellos en los que, por otro lado, la Administración de justicia encuentra su fundamental razón de ser) no son una simple manifestación de un desencuentro entre las partes, sino que canalizan, de forma más o menos manifiesta, cuestionamientos sobre el fundamento de la norma jurídica. Más que plantear qué cabe hacer conforme a Derecho, este tipo de situaciones contribuyen a reactivar una discusión sobre lo que es Derecho justo, siquiera a pequeña escala.

Sin embargo, no es necesario pensar en la actividad del Tribunal Europeo de Derechos Humanos para encontrar rasgos de lo que aquí se trata de canalizar, pues toda decisión jurisdiccional comparte elementos comunes que derivan de su función de «hacer justicia». Hacer justicia implica, en este sentido, algo más que *adjudicar*, pues supone un nuevo pronunciamiento sobre aquello que las personas *deben* hacer. Este deber se entiende mejor no desde un punto de vista meramente legalista, sino desde la perspectiva más profunda de la *corrección* axiológica.

La idea de corrección está fuertemente vinculada a la argumentación jurídica. Cuando el juez ofrece razones para dictar su fallo, el propósito último de aquellas es *justificar* su posición, es decir, hacerla aparecer como una descripción de la pauta de acción adecuada conforme a un correcto entendimiento del Derecho y sus valores. Esto no tiene por qué suceder de modo directo o completo: puede ser suficiente argumentar por qué la parte a la que se retira la razón actuó de manera incorrecta. Sin embargo, aquello que es común es la «no trivialidad» de las decisiones judiciales: detrás de un análisis del ajuste entre la conducta y el Derecho subyace un mensaje sobre aquello que está bien o mal hacer. Ambas dimensiones se comunican siempre y cuando se opera (o, como mínimo, el juez asume que opera) dentro de los marcos de un Derecho razonablemente justo que incorpora en su seno los cánones básicos de moralidad racional susceptibles de ser compartidos por todos (Atienza, 2013: 560-4).

En un sentido al menos ideal, la argumentación jurídica no solo pretende contribuir a ofrecer la interpretación más correcta posible del asunto (Taruffo, 1998:312), sino que cumple también la función de comunicar el Derecho aplicable, de tal manera que, si las partes intervinientes fueran racionales, habrían de quedar convencidas[7]. El elemento de persuasión racional/convencimiento (Atienza, 2013: 414; y García Amado, 2016:58) implica cierto compromiso de los intervinientes —y, particularmente, del juez— con el dis-

curso enunciado y con la práctica social en la que están insertos. Asimismo, contribuye a la labor de legitimar el sistema jurídico en su conjunto a través del caso concreto.

Cuando extrapolamos esta reflexión a la hipótesis del «juez IA», sin embargo, surgen serias dudas sobre la capacidad de un algoritmo (incluso tan «inteligente» como el que se imaginaba *supra*) para participar de este proceso comunicativo que, en principio, es característicamente social.

La cuestión de la pretensión de corrección es problemática, en la medida en que remite a un *compromiso* del actor con su discurso y otorga a este último ese carácter «no trivial». Describir de este modo la actividad jurisdiccional resulta muy similar a hablar de estados mentales y, en tal caso, nos enfrentamos de nuevo a la objeción de Searle. La única forma de armonizar «pretensión de corrección» y decisión algorítmica[8] sería retrotraernos al momento del diseño del algoritmo (obviando por un segundo su capacidad para reconfigurarse) y tratar de ver en las decisiones técnicas tomadas la intención de producir un protocolo de decisión tan correcto como fuera posible. No parece, sin embargo, que estemos hablando de lo mismo: la elaboración de un sistema que resulte adecuado en la mayoría de los casos es bastante más similar a la labor de legislar que a la de aplicar el Derecho. Asimismo, parece indispensable que el compromiso con una determinada visión del Derecho aplicable proceda de un agente al que podamos considerar responsable.

Es cierto, por otro lado, que «pretender corrección» y alcanzarla son cosas diferentes. Una decisión puede ser correcta sin pretenderlo (y a la inversa) y no es indispensable que el propio agente desee llegar subjetivamente a la respuesta correcta. Sin embargo, centrar la atención en la corrección del *resultado* priva a la decisión de cierto elemento «humano» que parece imprescindible (Sourdin, 2018: 1124). Con todo, la posibilidad de que los sistemas automatizados de decisión puedan ser más acertados y objetivos porque están desprovistos de las limitaciones cognitivas y sesgos ideológicos de las personas sigue siendo un argumento de peso, aunque tal vez algo optimista a la luz de los datos disponibles (Citron y Pasquale, 2014: 4; Burrell, 2016: 3; y Hannah-Moffat, 2018: 7). Con todo, si el desarrollo técnico permitiera en un futuro materializar las aspiraciones de los más entusiastas defensores de la IA, cabe plantearse cuánto tardaría en sacrificarse el aspecto «artesanal» que atribuimos a la labor de juzgar en favor de una mayor «eficiencia» y «objetividad».

Esto nos conduce, de nuevo, a la cuestión del convencimiento intersubjetivo y de la función social de los procesos de justicia. La tarea de «hacer justicia» se ha caracterizado como un espacio para el intercambio de pretensiones de legitimidad (sobre la conducta y la norma) entre los intervinientes. El prototipo del juicio justo y legítimo (y de la argumentación adecuada) sería aquel que el propio culpable estuviera dispuesto a aceptar, si fuera racional, en virtud de los motivos aducidos en la justificación. Pero ¿es solo una cuestión de razones (o, incluso, de buenas razones)? ¿O es relevante también la procedencia de estas? Si lo único relevante son los motivos de la decisión, presentados bajo la forma de argumentos, entonces una IA que pasara el «test de Toulmin» podría realizar satisfactoriamente esta labor por mucho que, como hemos visto, su «inteligencia» no fuese más que una simulación conductual del pensamiento humano.

Sin embargo, resulta pertinente preguntarse si la actividad jurisdiccional no está pensada para ser, al menos, preponderantemente humana. El proceso judicial, como espacio de encuentro de sujetos responsables con visiones contrapuestas sobre un conflicto, resulta especialmente propicio para establecer un marco de mínimo respeto entre los interlocutores en cuanto que participantes válidos de un proceso de intercambio de significados. Preservar este tipo de espacios y prácticas puede no ser especialmente eficiente o, incluso, demasiado fructífero; sin embargo, la necesidad parece proceder en última instancia de cierta idea de dignidad humana (Sourdin, 2018: 1127-30). Una forma de dignidad que tal vez requiere que la actividad de hacer justicia siga siendo una actividad *intersubjetiva* entre *pares*, y que deriva de la idea de que los asuntos humanos (importantes) solo pueden ser resueltos por humanos.

5. ALGORITMOS COMO APOYO A LA DECISIÓN JUDICIAL. LOS VACÍOS DE ARGUMENTACIÓN

Hasta aquí se ha tratado de ofrecer una panorámica general de algunos aspectos relevantes de la discusión «IA y Derecho», así como una breve reflexión sobre la hipótesis del «juez IA». El presente apartado adopta un punto de vista distinto y más apegado a la práctica, dentro de los cánones de lo que resulta concebible hacer con un algoritmo inteligente en la actualidad.

Con independencia de la clase de futurismo que se esté dispuesto a aceptar en relación con el desarrollo de la inteligencia artificial y sus eventuales repercusiones, resulta indiscutible que este tipo de instrumentos algorítmicos ya ha comenzado a hacerse un espacio como *complemento* para la toma de decisiones de jueces y tribunales (Martínez Garay, 2018: 488). Esta manifestación más cotidiana de IA, sin embargo, también presenta sus propios problemas, algunos de los cuales son reconducibles a una perspectiva argumentativa como la aquí abordada. En este sentido, la cuestión a dilucidar ya no es si la labor del juez puede ser suplantada por un algoritmo, sino de qué manera el empleo de estos últimos como *apoyo* puede transformar, bajo ciertas condiciones, la actividad jurisdiccional y, en particular, el papel argumentativo del discurso del juez.

No obstante, con el objetivo de reconducir esta parte de la discusión a un plano más concreto, se discutirá específicamente el uso de la inteligencia artificial en la realización de pronósticos de riesgo de reiteración delictiva[9] cuando estos son utilizados como elementos de juicio por los jueces o tribunales sentenciadores. Tal vez este no sea el caso más usual o directamente aplicable dentro de nuestras coordenadas jurídicas, pero sus particularidades permiten presentar con especial claridad algunos aspectos conflictivos del problema que nos ocupa.

5.1. LA DISCUSIÓN A TRAVÉS DE UN EJEMPLO: LA POLÉMICA DE *STATE VS. LOOMIS*

En la discusión anglosajona en materia de algoritmos y toma de decisiones judiciales, el caso *State of Winsconsin vs. Loomis* ha sido frecuentemente analizado a lo largo de los úl-

timos años[10]. Esto se debe, fundamentalmente, al carácter multidimensional de la controversia dirimida en el mismo, que prácticamente no deja aspecto sin tratar en relación con el uso de algoritmos en la impartición de justicia. Muy someramente, las líneas maestras del asunto son las siguientes.

En el contexto de un proceso penal por un tiroteo, en el que el Sr. Loomis afirmó no haber tomado parte más que como conductor de uno de los vehículos vinculados al mismo, se emitió un informe que incluía una valoración del riesgo de reiteración delictiva a través de la herramienta COMPAS.

COMPAS es un algoritmo de valoración del riesgo, propiedad de la empresa Equivant (Northpointe anteriormente), que se encarga de evaluar el riesgo que presenta un individuo en tres dimensiones: riesgo vinculado a la liberación provisional, riesgo de reincidencia general y riesgo de reincidencia violenta. En este caso, Loomis puntuaba alto en las tres dimensiones, lo que tuvo repercusiones en su sentencia[11], así como en una ulterior denegación del acceso a la libertad condicional.

En consecuencia, Loomis recurrió el uso que se había hecho del algoritmo COMPAS de acuerdo con tres premisas: *i)* que contravenía su derecho a ser juzgado conforme a información adecuada, en la medida en que el carácter privado del algoritmo bloqueaba la posibilidad de examinar su validez; *ii)* que el empleo de este instrumento conculcaba su derecho a una sentencia individualizada (por el carácter estadístico/grupal de los datos en que se basan tales instrumentos); y *iii)* por el efecto discriminatorio de tomar en consideración la variable de sexo, con efectos negativos en términos de predicción del riesgo (Wisser, 2019: 1813-6). En suma, lo que el caso *Loomis* planteaba era la posibilidad de *discutir* las predicciones de un algoritmo (Martínez Garay, 2018: 491; y Washington, 2019: 4).

La controversia en torno a COMPAS en particular, y al uso de instrumentos de valoración del riesgo de un tipo u otro en general, ha sido objeto de un creciente interés a partir de la publicación de una serie de informes elaborados por el grupo ProPublica en los que se denunciaba un sesgo racial implícito en el algoritmo (Angwin *et al.*, 2016; Dieterich, Mendoza y Brennan, 2016; Jeff Larson *et al.*, 2016; y Rudin, Wang y Coker, 2018). Aquí, sin embargo, nos interesan únicamente los aspectos del caso que resultan pertinentes para la discusión sobre algoritmos e inteligencia artificial.

5.2. LA CUESTIÓN DE LA OPACIDAD

La polémica sobre COMPAS ha sido prácticamente insoslayable en el debate actual, por mucho que buena parte de los problemas que acompañan a los protocolos de valoración del riesgo son comunes, sean o no resultado de técnicas de *machine learning*. Entre estas dificultades, aquí interesa abordar el problema de la *opacidad* y, en particular, la cuestión de si los algoritmos «inteligentes» son especialmente proclives a funcionar como «cajas negras» inaccesibles al escrutinio externo (Leese, 2014; Pasquale, 2015; Edwards y Veale, 2018; Rudin, Wang y Coker, 2018; y Wisser, 2019).

De acuerdo con Burrell (2016: 3-5), la opacidad de los algoritmos de *machine learning* puede adoptar tres formas distintas (a la vez, o de manera separada). En primer lugar, tenemos la opacidad derivada del secreto público o privado, que impide el escrutinio interno del algoritmo. Este es el tipo de opacidad que más claramente se manifestaba en el caso *Loomis* y es la base de la explotación comercial de instrumentos de este tipo (Martínez Garay, 2018: 491-2; y Rudin, 2018: 5-6). En segundo lugar, tenemos la opacidad derivada de la falta de conocimientos técnicos sobre el funcionamiento de un algoritmo o la interpretación de código informático. Dado que este campo de conocimiento está todavía reservado a un sector reducido y crecientemente especializado de la población, tratar con algoritmos y programas complejos, en general, resulta esotérico para la mayoría de personas, entre ellas los juristas (Martínez Garay, 2018: 495; y Washington, 2019). Esto resulta especialmente relevante cuando se trata de valorar si el uso de determinados procedimientos automatizados resulta, o no, conforme a las reglas y principios de un ordenamiento jurídico, y tiene bastante que ver con la controversia en torno al informe de ProPublica. Por último, encontramos un tercer tipo de opacidad relativo al propio funcionamiento de los algoritmos de *machine learning*, cuya peculiaridad estriba en el hecho de que incluso aquellos sujetos que poseen los conocimientos técnicos requeridos pueden llegar a ser incapaces de interpretar los motivos por los que el programa produce un determinado resultado. Esto último, a su vez, se vincula al menos a dos circunstancias[12]. Por un lado, desde el punto de vista de la *complejidad*, en un algoritmo de esta clase convergen un código complicado que opera a escala de *Big Data*, buscando patrones y correlaciones y alterándose a sí mismo a medida que «aprende». Por otro, desde el punto de vista de la *interpretación*, hemos visto cómo el modo de proceder de un algoritmo de estas características puede ser, en ocasiones, irreconducible a una lectura satisfactoria desde un punto de vista *semántico*[13].

5.3. VACÍOS ARGUMENTATIVOS

Aunque buena parte de las fuentes de donde proceden las dificultades que ahora examinaremos son similares a las que se han señalado al discutir la hipótesis del «juez IA», la cuestión a analizar es diferente. ¿Cómo reconducir el empleo de algoritmos inteligentes a la labor argumentativa del juez, especialmente en aquellos casos en que estos instrumentos desarrollan un papel esencial en el contenido o la dirección del razonamiento?

Pensemos de nuevo en la función de las valoraciones del riesgo. Las aplicaciones prácticas de este tipo de herramientas son múltiples: desde las medidas cautelares a la ejecución penitenciaria, pasando por la valoración en sede judicial, que es la que nos interesa. En este último ámbito, el algoritmo trata de recopilar toda la información relevante que ha resultado predictiva en casos análogos al que se somete a estudio con el fin de emitir un pronóstico sobre el riesgo que un individuo representa en términos de reincidencia delictiva (*in extenso*, Martínez Garay, 2014 y 2016; y Martínez Garay y Montes Suay, 2018). De ello se infiere que ni el núcleo del procedimiento ni sus objetivos son algo nuevo o ajeno a la justicia penal, sino más bien un renovado intento de sistematizar o racionalizar espacios de decisión anteriormente canalizados por otra vía (la intuición del juzgador, el peritaje clínico…).

Se trata, por lo tanto, de espacios de toma de decisión que, de no existir los protocolos, serían reconducidos a una valoración personal en un contexto de acusada incertidumbre. El problema de la toma de decisiones en tales contextos es que se reduce la solidez de los criterios sobre los que adoptar una posición razonable y, con ello, se incrementa el riesgo de arbitrariedad (o, al menos, de ilegitimidad). Un mecanismo para mantener tales riesgos bajo control proviene, habitualmente, de la elaboración de juicios suficientemente razonados y completos. La premisa de que, en condiciones normales, los acontecimientos futuros proceden de los pasados y están sometidos a las mismas leyes naturales y principios de racionalidad es una forma habitual de extrapolar una expectativa del presente al futuro. La clave de tales pronósticos es que, asentados en premisas sólidas y aparentemente aceptables, reducen la impresión de incertidumbre y reconducen la decisión adoptada dentro del rango de lo aceptable.

Transformar lo arbitrario en razonable, en lo que a pronosticar el futuro se refiere, obliga a las personas a realizar una importante labor argumentativa. En ella, el paso de las premisas a las conclusiones, la estructuración de líneas argumentales, la ponderación de distintas hipótesis alternativas, etc. se orienta en su totalidad a construir esta expectativa aceptable sobre un futuro incierto. En el contexto en que nos encontramos (el de las decisiones judiciales), además, esta expectativa se inserta dentro de un procedimiento argumentativo más amplio como un eslabón de la cadena que conduce al fallo. La dificultad de los pronósticos de riesgo dentro de los razonamientos construidos por el juez penal estriba en que, en ocasiones, los eslabones se contaminan entre sí de forma poco clara. De este modo, la percepción sobre el riesgo afecta al juicio de culpabilidad y viceversa, de modo que algunos elementos de juicio adquieren valores opuestos desde unas coordenadas u otras (Monahan y Skeem, 2015: 504).

Sin embargo, lo que importa aquí es una cuestión distinta. Lo trascendente es, de nuevo, que realizar un pronóstico sobre el futuro implica una labor eminentemente argumentativa, especialmente cuando se lleva a cabo sin el apoyo de instrumentos técnicos. El uso de estos últimos pretende, precisamente, reducir este espacio de discrecionalidad estandarizando el procedimiento y, en consecuencia, sustituir el esfuerzo argumentativo por el *output* de la herramienta. Sin embargo, el problema de la opacidad, cuando se integra dentro de este contexto, es que propicia que asomen *vacíos* dentro del procedimiento argumentativo. Esto es, permite que las razones que conectan un paso argumentativo con el siguiente puedan permanecer inaccesibles.

Ello obliga a establecer mecanismos compensatorios para sostener, desde fuera, que el «punto ciego» introducido por un algoritmo opaco cumple, efectivamente, el papel argumentativo que se le presume. Se trata, por tanto, de sostener la racionalidad interna y la adecuación de los resultados ofrecidos por el algoritmo a través de *respaldos,* tomando prestadas las categorías de Toulmin (2003). A título de ejemplo, algunas variantes del argumento de autoridad (de un científico determinado, o de la *ciencia* en su conjunto) u otro tipo de inferencias «circundantes» pueden cumplir este papel. Su función argumentativa, en tales casos, es preservar la integridad del paso premisa-conclusión (y, en última instancia, la del proceso argumentativo en su conjunto) transformando en axioma la inferencia

que los conecta. Pero un pensamiento axiomático puede ser un pensamiento dogmático, y un pensamiento dogmático tiene mucho de arbitrario en cuanto determina la inaccesibilidad de algunas premisas al razonamiento crítico. De modo que, en una paradójica circularidad, los esfuerzos por huir de la arbitrariedad de quien decide ante la incertidumbre desembocan en la arbitrariedad de blindar las partes opacas que sostienen la decisión.

Desde luego, nada de esto constituye una valoración de la eficacia de unos métodos u otros para realizar determinados pronósticos. Si algo ha provocado que este tipo de protocolos aparezca como más fiable que los razonamientos del juzgador ha sido, en general, su mayor consistencia y eficacia predictiva (Berk, 2018: 7). Sin embargo, no queda claro que el paso de lo útil a lo razonable sea algo autoevidente: si un algoritmo efectuara pronósticos altamente acertados, pero no se tuviera la menor idea de cómo llega a tales conclusiones, ¿bastaría alegar su eficacia para justificar las decisiones tomadas? A nivel puramente especulativo (y en términos abiertamente provocadores), ¿sería admisible una «teología» secular de la inteligencia artificial?

Extraer algún tipo de «semántica» de la «sintaxis», sin embargo, no es necesariamente una labor destinada al fracaso, como ilustra el debate en torno a la admisibilidad de ciertas variables que correlacionan con el riesgo de reincidencia (*v. gr.* Harcourt, 2010; y Starr, 2014). Discutir sobre si es admisible que un algoritmo tenga en cuenta el sexo o la raza en sus cálculos implica imponer, desde el plano de los significados, una determinada forma a los instrumentos y a las operaciones que es legítimo efectuar. Decir que un determinado cruce de variables resulta discriminatorio es un ejercicio interpretativo que no viene frenado por el carácter formal de un análisis correlacional.

Además, recientemente se ha puesto en duda la necesidad misma de acudir a modelos opacos de *machine learning* para obtener resultados adecuados. De este modo, antes que tratar de interpretar arduamente algoritmos herméticos, bastaría con hacer uso de instrumentos interpretables por diseño cuando ello no redunde en pérdidas de eficacia notables (Zeng, Ustun y Rudin, 2017; Rudin, 2018; Rudin, Wang y Coker, 2018; y Rudin y Ustun, 2019).

5.4. ALGORITMOS OPACOS Y JUECES OPACOS. SOBRE LOS CONTEXTOS DE DESCUBRIMIENTO Y JUSTIFICACIÓN

Llegados a este punto, cabe hacer una breve reflexión sobre la idea de opacidad y el modo en que la aplicamos a unos casos u otros. Se ha dicho que, por diversos motivos, algunos algoritmos resultan opacos. Significativamente, algunos lo son por el propio procedimiento que siguen para detectar los patrones en los datos que darán lugar a las ulteriores clasificaciones. Esto, sin embargo, no es tan extraño como puede parecer; cabría incluso afirmar que es una forma característicamente humana de proceder: la intuición tiende a preceder a la razón. Mas la intuición es tan útil como inescrutable en la mayoría de los casos, y así es como viene siendo aceptada. El recurso a las intuiciones, las corazonadas y las sensaciones supone, habitualmente, el punto final de la explicación, el momento a partir del cual no es posible remontar.

Esto es lo que, siguiendo a Reichenbach (1938), se ha incorporado a la idea de *contexto de descubrimiento*, es decir, el conjunto de factores psicológicos, biográficos, culturales, coyunturales, etc. que permiten explicar, desde un punto de vista externo, la aparición de una

idea. Se trata del *contexto* dentro del cual llega a producirse el *descubrimiento* por parte del agente y, como tal, resulta también particularmente magro a la hora de proporcionar *razones*. Este es, de hecho, el elemento característicamente problemático, dado que ninguna de las circunstancias que hacen surgir una idea están semánticamente conectadas con la idea en sí. En el caso de las decisiones, nada del contexto de descubrimiento permite, *a priori*, sostener como razonable la elección hecha. En relación con la cuestión objeto de estudio, una corazonada resulta tan ilustrativa como la detección de patrones de una *neural network* en materia de IA y, en este sentido, permite retratar al juez como «mecanismo opaco de toma de decisiones» (Burrell, 2016: 7).

Salir del aparente nihilismo proyectado por este (someramente retratado) callejón sin salida irracionalista explica la necesidad de acudir a la segunda parte de la dicotomía de Reichenbach, *el contexto de justificación.* En este segundo momento, se trata de reconducir la idea «emergida» a un esquema de racionalidad preexistente y en el que tiene que pasar a ocupar un papel. De este modo, una hipótesis científica, por ejemplo, debe mostrar que es aceptable desde los cánones científicos establecidos, insertándose en el *corpus* preexistente que contribuye a desarrollar. La labor principal es, por tanto, *justificativa*, lo que implica defender que la relación entre la hipótesis y el resto del cuerpo de conocimientos es genuina, una relación de la parte al todo. Esta es una empresa típicamente racional que otorga primacía a la labor reflexiva del agente, que es la que permite «emancipar» a la idea respecto de su contexto de descubrimiento.

En el caso del Derecho y los jueces, la argumentación jurídica desempeña precisamente este papel justificativo, pues insertar una idea (ya sea una interpretación, ya una decisión) dentro del marco jurídico en el que ha de intervenir y con el que debe armonizar. El proceso racional llevado a cabo para justificar la idea original según los parámetros de legitimidad del Derecho supone, en cierto modo, una *garantía* sobre la adecuación de las conclusiones, en la medida en que permite, como mínimo, filtrar la arbitrariedad palmaria embridando la argumentación dentro de restricciones de admisibilidad que condicionan su forma y su fondo. Así, el razonamiento justificativo es el mecanismo que permite trazar una unión en la «dicotomía de contextos» y, con ello, realizar la conversión de lo «irracional» en «racional».

Pero si el problema de la opacidad del juez se resuelve[14] enfatizando que el valor de sus decisiones deriva del marco de racionalidad en el que son reconstruidas y elaboradas (argumentativamente), la cuestión es: ¿cabe efectuar una operación análoga con los algoritmos? Aunque seguramente no sea posible ofrecer una respuesta en un trabajo de estas características, tal vez sea pertinente recurrir a una distinción que, en principio, parece desarmar la analogía.

Si la totalidad de las computaciones efectuadas por el algoritmo es opaca y solo tenemos acceso al *input* (variables predictivas) y al *output* (riesgo de reincidencia), carecemos de una fase intermedia de carácter argumental que permita una lectura del resultado en términos de significado. Dicho de otro modo, sería como una sentencia escrita sobre la base de la intuición (opaca) que el juez hace tras examinar las circunstancias del caso. Además, falta un marco de racionalidad preestablecido sobre el que insertar el pronóstico de riesgo de manera coherente y, así, contribuir a sostenerlo. Algo parecido podría hacerse mediante la reconducción de las bases estadísticas sobre las que se asienta el algoritmo a la teoría estadística general, o de las variables de riesgo empleadas a la teoría criminológica correspon-

diente, si bien no queda claro que esta opción resulte plenamente satisfactoria. El problema fundamental parece residir, precisamente, en el hecho de que en materia de algoritmos tal vez no sea posible diferenciar entre descubrimiento y justificación.

5.5. DISCRECIONALIDAD Y LA BUROCRATIZACIÓN DE LA DECISIÓN DISCRECIONAL

Por último, y a modo de cierre, cabe mencionar una última forma en que el uso de algoritmos puede repercutir en la toma de decisiones judiciales que, si bien no procede directamente del problema de la opacidad, puede ser agravada por ella. Se trata, pues, de examinar el modo en que el uso de protocolos y algoritmos puede afectar a la decisión discrecional.

Los espacios discrecionales de decisión judicial han sido tradicionalmente problemáticos para la teoría del Derecho y han acaparado buena parte de la controversia en torno al correcto funcionamiento del ordenamiento jurídico. Ello se entiende especialmente bien a la vista de que, tras la discusión sobre los espacios de discrecionalidad judicial, subyace el problema de la correcta articulación de la separación de poderes.

Desde un punto de vista funcionalista puede entenderse, por otro lado, que existe una suerte de diálogo entre el poder legislativo y el judicial. Los espacios discrecionales, derivados de la «textura abierta» del Derecho (Hart, 1963: 159), de la necesidad de interpretar para solventar cuestiones de vaguedad o ambigüedad o, incluso, de expresas remisiones previstas por el legislador, cumplen su función en la medida en que, por un lado, evitan la elaboración de una ley sumamente pormenorizada que adolezca de excesiva rigidez y, por otro, habilitan al juez a adaptar la norma jurídica al caso concreto. Se trata, pues, de un mecanismo de transición de lo general a lo particular que tan solo puede desarrollarse eficazmente en la medida en que presupone un juez obligado a *justificar* con especial intensidad aquellas decisiones tomadas conforme a sus facultades discrecionales. En el fondo, se trata de discrecionalidad *porque* es *limitada*[15], es decir, circunscrita siempre a los marcos generales de lo jurídicamente aceptable (si no vía reglas, al menos a través de principios).

El recurso a la discrecionalidad judicial (bien entendida) es, por tanto, un mecanismo legitimador del ordenamiento jurídico: por una parte, porque permite adaptar la ley al caso concreto; por otra, porque se trata de un elemento central para dotar al Derecho de un grado suficiente de dinamismo y adaptabilidad sin sacrificar cotas excesivas de seguridad jurídica. En este sentido, la «complejidad» y la relativa indeterminación del marco normativo devienen funcionales cuando se contrapesan con una obligación justificativa en sentido fuerte.

Sin embargo, los espacios discrecionales siempre han resultado problemáticos, y sería equivocado entender, como resultado de esta representación un tanto consensual del estado de cosas, que no existe un sector importante de opinión que preferiría reducir la indeterminación e imprevisibilidad que conlleva la discrecionalidad. En este sentido, *racionalizar* la discrecionalidad sigue siendo el terreno sobre el que se tratan de formalizar, mediante unos mecanismos u otros, los criterios a emplear por el juzgador para hacer un uso adecuado y, precisamente, «racional» del poder que ostenta.

Ejemplos recientes de esto pueden encontrarse, especialmente en el ámbito anglosajón, en las *Sentencing Guidelines* (*v. gr.* Nagel, 1990), el movimiento por el *evidence-based sentencing* (Wolff, 2008; Warren, 2009; Oleson, 2011; y Starr, 2014) o, en lo que aquí interesa, el uso de algoritmos para apoyar la decisión del juez[16]. El uso de algoritmos de apoyo puede, así, adoptar diversas formas: desde efectuar un cribado de las posibles «opciones» que se abren al juzgador a la vista de las circunstancias del caso hasta introducir elementos de juicio sobre aspectos discrecionales como, precisamente, el riesgo.

La duda que surge a raíz de las nuevas potencialidades brindadas por el desarrollo de los algoritmos estriba en determinar de qué manera podrían afectar al desempeño por parte de los jueces de sus facultades discrecionales. En este sentido, existe un riesgo en todo proceso de simplificación y estandarización: la pérdida de grados de complejidad que pudieran resultar necesarios para el correcto desempeño de la labor correspondiente (Taruffo, 1998: 319). Asimismo, parece que la propia idea de estandarización conlleva una reducción de la variabilidad, cuya valoración positiva o negativa estará fuertemente influida por el peso que se atribuya a la seguridad jurídica frente a la «individualización» judicial.

En este sentido, se ha criticado que la *racionalización* de las decisiones discrecionales puede mermar la discrecionalidad misma precisamente por la influencia *burocratizante* de este uso de los algoritmos (Taruffo, 1998: 321-322; y Peeters y Schuilenburg, 2018: 270). Y aunque es cierto que no hay motivos para pensar que el juzgador no disponga de la facultad para desoír el resultado del algoritmo si lo considera inadecuado al caso concreto, parece que la forma en que la automatización de las decisiones influye en los agentes decisores justifica cierto escepticismo al respecto (Goddard, Roudsari y Wyatt, 2012; y Washington, 2019:35). Frente a la «seguridad» que ofrece seguir el criterio ofrecido por el algoritmo, cualquier distanciamiento supone un riesgo y, con él, la consecuente necesidad de realizar esfuerzos argumentativos adicionales para sofocar la sospecha de arbitrariedad. En el campo de la valoración del riesgo de reincidencia delictiva, resulta difícil esperar algo distinto a una inclinación securitaria (Peeters y Schuilenburg, 2018: 274-5).

Si los usuarios de un protocolo acaban abundando, generalmente, en la dirección señalada por el mismo, la tendencia emergente apunta hacia una proclividad sistémica a reproducir los criterios y valores ínsitos en su diseño, una situación potencialmente problemática desde el punto de vista de la distribución del poder en los sistemas democráticos (Sourdin, 2018:1126). Y es que, a pesar de su apariencia técnica y objetiva, la elaboración de un instrumento de valoración de riesgo (siguiendo con el caso escogido para orientar la discusión) implica una multiplicidad de decisiones de ajuste, selección y ponderación de variables, criterios de discriminación entre riesgos bajos, medios o altos, así como tasas de errores (falsos positivos y negativos) tolerables (Martínez Garay, 2016; Berk, 2018: 10-3). Todas estas pequeñas decisiones son, en realidad, cuestiones ético-políticas *sustraídas* de su campo legítimo y *ocultadas* tras el velo de la técnica. Desde un punto de vista argumentativo, nuevamente, nos encontramos ante todo un campo de cuestiones problemáticas que serían objeto de atención, discusión y justificación si no hubieran quedado externalizadas, generando, de nuevo, un relevante vacío argumentativo.

Si, como frecuentemente afirman sus defensores, el uso de este tipo de algoritmos supone un modo de limitar los sesgos implícitos que se deslizan en el ordenamiento jurídico a través de la discrecionalidad judicial, esto solo puede suponer un mérito en la medida en que el algoritmo no permita el mismo deslizamiento (Citron y Pasquale, 2014: 4; y Peeters y Schuilenburg, 2018: 274). Y si, como hemos visto, la única forma de paliar los riesgos de la «opacidad» del juez es el mandato de justificación de las decisiones, controlar los riesgos inherentes al uso de algoritmos exige necesariamente la *transparencia* de los mismos en los tres niveles anteriormente señalados. El carácter público y abierto del código permite a cualquier actor público o privado someterlo a prueba a través de múltiples bases de datos, contribuyendo a desentrañar más eficazmente las virtudes y vicios del algoritmo (Washington, 2019: 23). Adicionalmente, cuando la tarea lo permita, los algoritmos simples son preferibles a los complejos si se sitúan dentro de un rango semejante de eficacia (Rudin, 2018: 5). Por último, y de manera fundamental, parece razonable la propuesta de que, en campos de decisión de alta trascendencia no deberían emplearse algoritmos que funcionen como «cajas negras», especialmente cuando otros de carácter transparente pueden cumplir la tarea (Rudin, 2018: 7).

En una nueva oscilación de lo informático a lo jurídico, estas y otras medidas podrían contribuir a garantizar una suerte de *debido proceso tecnológico* (Citron, 2007; y Citron y Pasquale, 2014). A su vez, permitiría recuperar como materia de argumentación toda una serie de circunstancias, decisiones y procesos que, por «naturaleza», requerirían de una mínima justificación si no se hallaran ocultos tras la técnica. Por último, contribuiría a evitar que, por inadvertencia, los algoritmos en materia de pronóstico acabaran funcionando como profecías autocumplidas, confirmando con cada nueva aplicación las variables incorporadas en su diseño (Harcourt, 2010; y Citron y Pasquale, 2014:18)

6. CONCLUSIONES

A lo largo de los últimos años, el ámbito de la inteligencia artificial se ha caracterizado por su gran dinamismo. En buena medida, ello se debe a la rápida expansión de sus campos de aplicación práctica, propiciada por el enorme volumen de datos que las sociedades contemporáneas producen cotidianamente.

La esfera del Derecho ha sido, tradicionalmente, un foco de interés para la investigación en materia de inteligencia artificial. Este hecho ha estado, y sigue estando, fuertemente relacionado con el tipo de decisiones complejas que caracterizan a buena parte de la práctica jurídica. La capacidad de los sistemas algorítmicos más sofisticados para replicar, complementar y, en ocasiones, suplantar a los decisores humanos, sin embargo, ha sido también objeto de controversia cuando el debate ha tomado como referencia la labor del juez.

En este trabajo se han examinado algunos aspectos problemáticos de esta interacción disciplinaria. En concreto, lo que hemos denominado la hipótesis del «juez IA», y el uso de algoritmos para asistir o complementar al juez tradicional.

En relación con la hipótesis del «juez IA», las características de los algoritmos en materia de *machine learning* e IA son susceptibles de producir interesantes aproximaciones o predicciones sobre la conducta de jueces y tribunales. Sin embargo, no parece que puedan sustituir plena y completamente la decisión humana más que en aspectos relativamente triviales. La complejidad de las decisiones tomadas y los recursos de razonamiento empleados para resolver,especialmente, los denominados casos difíciles no parecen estar aún al alcance de los procedimientos computacionales. Asimismo, el modo en que «piensa» un algoritmo parece difícil de armonizar con una práctica que requiere de la producción de *razones* significativas. Finalmente, resulta dudoso que la sustitución del juzgador humano por el artificial pueda realizarse únicamente atendiendo a criterios de eficacia. La práctica de aplicar el Derecho parece impregnada de un elemento antropológico difícil de sacrificar y que obliga a plantearse si ciertas actividades sociales, como atribuir un reproche, solo tienen sentido cuando las realizan las personas.

En contraste con el carácter todavía fuertemente especulativo de la cuestión anterior, la posibilidad de emplear algoritmos como *apoyo* para el juzgador es una realidad que pugna por expandirse. Sin embargo, presenta una serie de dificultades vinculadas, especialmente, a la forma en que ciertos algoritmos funcionan como «cajas negras». Las dificultades para acceder al funcionamiento de aquellos o, simplemente, a interpretarlos contribuye a introducir puntos ciegos que se manifiestan como *vacíos de argumentación.* A diferencia de lo que ocurre en las decisiones judiciales, para los algoritmos no parece sencillo hallar un equivalente funcional del «contexto de justificación» que pueda compensar la falta de transparencia. Por último, el modo en que tiende a desarrollarse la interacción entre los decisores humanos y los protocolos automatizados reclama el empleo cauteloso de este tipo de «apoyos». Racionalizar la discrecionalidad judicial puede comportar el coste de reducirla o eliminarla, lo cual es una decisión de política jurídica que debe ser adoptada explícitamente antes que de manera inadvertida. Por otro lado, asumir la neutralidad axiológica y la objetividad de los algoritmos puede contribuir a legitimar decisiones ético-políticas tomadas en su diseño y que vienen legitimadas al cristalizar, de nuevo inadvertidamente, en el proceso. Estas dificultades no parecen ser inherentes al uso de algoritmos, «inteligentes» o no, sino a una concreta materialización de la relación entre Derecho e IA que ha permitido un uso acrítico de los desarrollos tecnológicos sin plena conciencia de sus limitaciones.

De acuerdo con los puntos examinados, por tanto, no parece descabellado hallar espacios para la penetración de la IA en la actividad jurisdiccional, especialmente cuanto menos ambiciosa (o más «instrumental») sea la tarea que se le otorgue. A la inversa, cuanto más entrelazados se encuentren este tipo de herramientas con el proceso decisorio, más imperativo resulta atender a eventuales obstáculos con el propósito de no desnaturalizar la labor de enjuiciamiento. Los «vacíos de argumentación» representan, precisamente, una de estas consecuencias indeseadas, en la medida en que toda una serie de decisiones y criterios problemáticos implícitos pueden quedar inadvertidamente sustraídos al escrutinio público, al tiempo que dan sustento, a su vez, a nuevas decisiones cuyo sustrato lógico o fáctico deja de ser totalmente transparente. Difícilmente se podría aquí ser categórico sobre si la IA resulta necesariamente nociva cuando entra en estrecha proximidad con procesos de toma de decisiones no triviales (aquí, el ámbito penal ha sido el ejemplo particular para ilustrar esta problemática

general). Más modestamente, sin embargo, la prudencia impone una actitud a medio camino entre el entusiasmo cauteloso y la sospecha paciente. Tal vez los algoritmos todavía tengan que demostrar que pueden convertirse en verdaderos *agentes*. Pero más intrigante es la duda de si las personas tendrán un mínimo interés por seguir siéndolo o si la posibilidad de deshacerse de la carga de la responsabilidad nietzscheana será demasiado tentadora.

NOTAS

1. Hablamos de «Beauty AI». Disponible en: <http://beauty.ai>. Se trata de unconcurso de belleza en el que la labor de juzgador la lleva a cabo un algoritmo tras haber sido «entrenado» con múltiples datos fisionómicos. Su estreno en 2016 vino acompañado inmediatamente por la polémica en torno a un sesgo racial imprevisto, dado que el algoritmo favorecía a concursantes con piel clara.

2. Aunque aquí se hace referencia, sobre todo, a los procedimientos técnicos para elaborar sistemas de inteligencia artificial, esta multiplicidad de procedimientos guarda relación con la falta de acuerdo sobre qué es lo característico de la inteligencia humana y qué modelo debe ser el que sirva como pauta para la IA. Desde este punto de vista, no es lo mismo pensar o actuar racionalmente que pensar o actuar *humanamente*, por mucho que la idea de racionalidad siga siendo central en la definición de lo humano (Russell and Norvig, 2010: 2-5).

3. En analogía con la expresión relativa a la teoría del Derecho, aunque salvando las distancias en cuanto a las implicaciones. Los trabajos de Dung (1995) y Prakken (1995) suelen destacarse como textos especialmente relevantes en la apertura la IA y el Derecho hacia una perspectiva argumentativa.

4. Merece la pena precisar, no obstante, el carácter más bien «teórico» de la exposición, cuyo objetivo es servir a la reflexión antes que plantear una posibilidad concreta y pragmática. Si fuera este el propósito, debería indicarse que las dos hipótesis examinadas no se encuentran en igualdad de condiciones, de modo que la posibilidad de sustituir completamente al juez (especialmente en materia penal, como aquí interesa) por un procedimiento automatizado de decisión a través de IA no resulta particularmente verosímil de acuerdo con la discusión internacional al respecto, o con la propia normativa nacional en materia de protección de datos. Con independencia de los detalles, el rechazo de fondo (o, al menos, la suspicacia) hacia una labor judicial automatizada, por no decir robotizada, hace que este aspecto sea relativamente menos debatido que la segunda hipótesis, a saber, el uso de la IA como instrumento al servicio de la Administración de justicia. Obsérvese al respecto la Carta ética europea sobre el uso de la inteligencia artificial en los sistemas judiciales y su entorno (2019) adoptada por la Comisión Europea para la Eficiencia de la Justicia (CEPEJ), o el art. 22 del Reglamento General de Protección de Datos. De hecho, incluso usos como el de la herramienta COMPAS son objeto de «las más extremas reservas» de acuerdo con la Carta (p. 66).

5. El procedimiento, someramente descrito, no resulta necesariamente el más adecuado para realizar este tipo de tareas, desde un punto de vista técnico. Su uso es meramente ilustrativo, con el fin de dar cuenta del modo en que una argumentación podría ser artificialmente simulada. Un modelo interesante de simulación de la toma de decisiones judicial puede encontrarse en Li *et al.* (2018).

6. Que aquí asociamos al esquema de Toulmin (2003) por ser uno de los más sencillos y, por tanto, ilustrativo de lo *indispensable* para una argumentación que pueda parecer razonable. Por lo demás, se trata de un modelo que ha resultado altamente influyente en los trabajos sobre IA y Derecho (Feteris, 2017: 55-9).

7. Algo que, por cierto, queda inmediatamente suprimido en un contexto en que tanto el «juez» como el fiscal o el abogado fueran sustituidos a su vez por inteligencias artificiales. Como bien recuerda Barona Vilar, no parece que tenga mucho sentido hablar de retórica, persuasión o convencimiento cuando todos los interlocutores son artificiales, al menos de acuerdo con el estado actual de los conocimientos (Barona Vilar, 2021: 576).

8. Desde luego, se podría defender que el requisito de la pretensión de corrección resulta superfluo y en absoluto imprescindible para desempeñar adecuadamente la tarea que viene encomendada a jueces y tribunales. Desde este punto de vista, no sería fundamentalmente distinto resolver una controversia jurídica en un juicio penal que resolver cualquier otra situación de carácter burocrático. La única diferencia procedería de variables extrasistémicas (importancia de la controversia para los afectados, repercusiones del conflicto, etc.) sin mayor repercusión desde el punto de vista interno (desde el punto de vista de aplicar el Derecho al caso concreto). Con todo, este no es el punto de vista que se asume aquí.

9. Desde luego, este no es el único uso concebible de uso de la IA como asistencia a la labor del juez, sino simplemente el punto de interés escogido para este texto. Una panorámica de los diversos usos posibles de este tipo de instrumentos se encuentra en Nieva Fenoll (2018).

10. Dentro de nuestro contexto académico, destaca el análisis que lleva a cabo Lucía Martínez Garay (2018) de esta misma sentencia y los problemas que plantea.

11. La influencia del pronóstico de riesgo puede haber sido considerable, teniendo en cuenta la diferencia entre el año de prisión que habían negociado ambas partes, y la ulterior sentencia de 7 años, recurrida ante la Wisconsin Supreme Court (Washington, 2019: 11).

12. En este punto resulta pertinente una aclaración sobre la opacidad técnica, que no se efectúa a cuerpo de texto en la medida en que no entorpece el argumento, pero que puede ser oportuno dejar asentada: los algoritmos de *machine learning* no son inherentemente opacos, como la propia discusión efectuada en el texto ilustra. Sin embargo, cierto tipo de sistemas de ML, como los que se basan en redes neuronales profundas (*Deep learning*) sí funcionan como cajas negras desde el punto de vista del escrutinio externo. Este matiz es importante, en la medida en que permite diferenciar y seleccionar, de ser necesario, sistemas de aprendizaje automático que sean transparentes por diseño (Marcus, 2018: 10-11;).

13. En relación con el problema de la interpretabilidad de los resultados obtenidos a través de algoritmos como los descritos, existe una saludable corriente de investigación y desarrollo dedicada específicamente a salvar este inconveniente, comúnmente caracterizada como «Explainable IA» o «Interpretable IA». A este respecto, puede ser útil al lector la consulta de Samek *et al.* (eds.) (2019). Puede interesar igualmente, de manera condensada, Zednik (2019: 4-5)

14. Aquí se ha optado por una descripción no problemática de la distinción entre contextos de descubrimiento y justificación, así como el modo en que trata de solventarse el problema en Derecho. No obstante, ni lo escrito da cuenta de todas las dimensiones del problema ni presupone considerar resueltas las críticas del realismo jurídico. Igualmente, tampoco implica aceptar que la dicotomía de Reichenbach deba tomarse al pie de la letra y que ambos contextos funcionen como compartimentos estancos. Su utilidad es, en este caso, ilustrativa, un modo de canalizar la discusión a grandes trazos, antes que adentrarse en sus pormenores.

15. En un sentido similar a lo que Taruffo (1998: 319) denomina discrecionalidad en sentido *débil* o *regulada*.

16. Parcialmente solapado con el movimiento *evidence-based* en la medida en que los algoritmos de predicción del riesgo, por ejemplo, son instrumentos «basados en evidencia» científica que, según sus partidarios, permitirían restringir los efectos perniciosos de la discrecionalidad judicial ofreciendo criterios sólidos sobre los que efectuar una decisión. Al respecto pueden consultarse los trabajos citados.

BIBLIOGRAFÍA

ALETRAS, Nikolaos, Dimitrios TSARAPATSANIS, Daniel PREOŢIUC-PIETRO y Vasileios LAMPOS (2016): «Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective», *PeerJ Computer Science*, 2, 1-19. doi: 10.7717/peerj-cs.93.

ANGWIN, Julia, Jeff LARSON, Surya MATTU y Lauren KIRCHNER (2016): «Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks», *ProPublica*.

ATIENZA, Manuel (2013): *Curso de argumentación jurídica*, Madrid: Trotta.

– (2017): «El juez perfecto», *Jueces para la democracia,* 90, 43-48.

BARONA VILAR, Silvia (2021): *Algoritmización del Derecho y de la Justicia. De la inteligencia artificial a la Smart Justice*, Valencia: Tirant lo Blanch.

BELLOSO MARTÍN, Nuria (2019): «Algoritmos predictivos al servicio de la justicia: ¿una nueva forma de minimizar el riesgo y la incertidumbre? », *Revista de la Faculdade Minera de Direito,* 22(43), 1-31.

BENCH-CAPON, Trevor J. M. y Paul E. DUNNE (2007): «Argumentation in Artificial Intelligence», *Artificial Intelligence*, 171(10–15), 619-641. doi: 10.1016/j.artint.2007.05.001.

BERK, Richard (2018): *Machine Learning Risk Assessments in Criminal Justice Settings*, *Machine Learning Risk Assessments in Criminal Justice Settings*, Nueva York: Springer. doi: 10.1007/978-3-030-02272-3.

BOYD, Danah y Kate CRAWFORD (2012): «Critical Questions for Big Data», *Information, Communication & Society*, 15(5), 662–679. doi: 10.1080/1369118x.2012.678878.

BURRELL, Jenna (2016): «How the machine "thinks": Understanding opacity in machine learning algorithms», *Big Data & Society*, 3(1), 1-12. doi: 10.1177/2053951715622512.

CEPEJ (2019): European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment, Consejo de Europa.

CITRON, Danielle K. (2007): «Technological due process», *Washington University Law Review*, 85(6), 1249-1313.

CITRON, Danielle K. y Frank PASQUALE (2014): «The scored society: Due process for automated predictions», *Washington Law Review*, 89(1), 1-33.

COHN, Gabe (2018): «AI Art at Christie's Sells for $432,500», *The New York Times*, 25 October [en línea] <https://www.nytimes.com/2018/10/25/arts/design/ai-art-sold-christies.html>.

DIETERICH, William, Christina MENDOZA y Tim BRENNAN (2016): *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*.

DUNG, Phan M. (1995): «Artificial Intelligence. On the acceptability role in nonmonotonic of arguments and its fundamental reasoning, logic programming and n-person games», *Artificial Intelligence*, 77, 321–357.

EDWARDS, Lilian y Michael VEALE (2017): «Slave to the Algorithm? Why a "Right to an Explaination" is Probably not the Remedy you are Looking For», *Duke Law & Technology Review*, 16(1), 18-84.

– (2018): «Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"? », *IEEE Security and Privacy*, 16(3), 46-54. doi: 10.1109/MSP.2018.2701152.

FETERIS, Eveline T. (2017): *Fundamentals of Legal Argumentation*, Nueva York: Springer. doi: 10.1007/978-94-024-1129-4.

GARCÍA AMADO, Juan A. (2016): «¿Para qué sirve la teoría de la argumentación jurídica?», *Teoría & Derecho. Revista de pensamiento jurídico,* 20, 42-63.

GODDARD, Kate, Abdul ROUDSARI y Jeremy C. WYATT (2012): «Automation bias: A systematic review of frequency, effect mediators, and mitigators», *Journal of the American Medical Informatics Association*, 19(1), 121-127. doi: 10.1136/amiajnl-2011-000089.

GROSAN, Crina y Ajith ABRAHAM (2011): *Intelligent Systems. A modern approach*. Berlin: Springer.

HANNAH-MOFFAT, Kelly (2018): «Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates», *Theoretical Criminology*, March 2018, 1-18. doi: 10.1177/1362480618763582.

HARCOURT, Bernard E. (2010): «Risk As a Proxy for Race», *Public Law & Legal Theory Working Paper*, 27(323), 237-244. doi: 10.1525/fsr.2015.27.4.237.
HART, Herbert L. A. (1963): *El Concepto de Derecho*, Buenos Aires: Abeledo-Perrot.
LARSON, Jeff, Surya MATTU, Lauren KIRCHNER y Julia ANGWIN (2016): «How We Analyzed the COMPAS Recidivism Algorithm», *ProPublica* [en línea], <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.
LEESE, Matthias (2014): «The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union», *Security Dialogue*, 45(5), 494-511. doi: 10.1177/0967010614544204.
LEVIN, Sam (2016): «A Beauty Contest was Judged by AI and the Robots didn't Like Dark Skin», *The Guardian*, September 8 [en línea] <https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people>.
LI, Jiajing, Guoying ZHANG, Longxue YU y Tao MENG (2018): «Research and Design on Cognitive Computing Framework for Predicting Judicial Decisions», *Journal of Signal Processing Systems*, 91(1). doi: 10.1007/s11265-018-1429-9.
MARCUS, Gary (2018): «Deep Learning: A Critical Appraisal», *ArXiv* abs/1801.00631
MARTÍNEZ GARAY, Lucía (2014): «La incertidumbre de los pronósticos de peligrosidad: consecuencias para la dogmática de las medidas de seguridad», *Indret: Revista para el Análisis del Derecho*, 2(2).
– (2016): «Errores conceptuales en la estimación de riesgo de reincidencia. La importancia de diferenciar sensibilidad y valor predictivo, y estimaciones de riesgo absolutas y relativas», *Revista Española de Investigación Criminológica*, 14(3), 11-31.
– (2018): «Peligrosidad, Algoritmos y Due Process: El caso State vs. Loomis», *Revista de Derecho Penal y Criminología*, 20, 485-502.
MARTÍNEZ GARAY, Lucía y Francisco Montes Suay (2018): «El uso de valoraciones del riesgo de violencia en Derecho Penal: algunas cautelas necesarias», *InDret*, 2(2018), 1-47.
MARTÍNEZ GARCÍA, Jesús I. (2019): «Inteligencia y derechos humanos en la sociedad digital», *Cuadernos Electrónicos de Filosofía del Derecho,* 40, 168-189.
MEDVEDEVA, Masha, Michel VOLS y Martijn WIELING (2019): «Using machine learning to predict decisions of the European Court of Human Rights», *Artificial Intelligence and Law*, Online fir, 1–30. doi: 10.1007/s10506-019-09255-y.
MIRÓ LLINARES, Fernando (2018): «Inteligencia Artificial y Justicia Penal: Más allá de los resultados lesivos causados por robots», *Revista de Derecho Penal y Criminologia*, 20, 87-130.
MONAHAN, John y Jennifer L. SKEEM (2015): «Risk Assessment in Criminal Sentencing», *Annual Review of Clinical Psychology*, 12(1), 489-513. doi: 10.1146/annurev-clinpsy-021815-092945.
NAGEL, Ilene H. (1990): «Structuring Sentencing Discretion: The New Federal Sentencing Guidelines», *Journal of Criminal Law and Criminology*, 80(4), 883-943.
NIEVA FENOLL, Jordi (2018): *Inteligencia artificial y proceso judicial*, Madrid: Marcial Pons.
OLESON, James C. (2011): «Risk in sentencing: constitutionally suspect variables and evidence-based sentencing», *SMU Law Review*, 64(4), 1329-1402.
PASQUALE, Frank (2015): *The Black Box Society. The Secret Algorithms That Control Money and Information*, Cambridge: Harvard University Press.
PEETERS, Rik y Marc SCHUILENBURG (2018): «Machine justice: Governing security through the bureaucracy of algorithms», *Information Polity*, 23(3), 267-280. doi: 10.3233/IP-180074.
PRAKKEN, Henry (1995): «From Logic to Dialectics in Legal Argument», en *Proc. 5th ICAIL*, 165–174.
REICHENBACH, Henry (1938): *Experience and Prediction. An analysis of the foundations and the structure of knowledge*, Chicago: University of Chicago Press.
RISSLAND, Edwina L., Kevin D. ASHLEY y Ronald P. LOUI (2003): «AI and Law: A fruitful synergy», *Artificial Intelligence*, 150, 1-15. doi: 10.1016/S0004-3702(03)00122-X.

RUDIN, Cynthia (2018): «Please Stop Explaining Black Box Models for High Stakes Decisions», en *Proceedings of NeurIPS 2018 Workshop on Critiquing and Correcting Trends in Machine Learning*, 1-15 [en línea] <http://arxiv.org/abs/1811.10154>.
RUDIN, Cynthia y Berk USTUN (2019): «Optimized Scoring Systems: Towards Trust in Machine Learning for Healthcare and Criminal Justice», *Interfaces*, 59.
RUDIN, Cynthia, Caroline WANG y Beau COKER (2018): «The age of secrecy and unfairness in recidivism prediction», 1-46 [en línea] <http://arxiv.org/abs/1811.00731>.
RUSSELL, Stuart y Peter NORVIG (^{3}2010): *Artificial Intelligence. A modern approach*, Upper Saddle River, NJ: Pearson. doi: 10.1017/S0269888900007724.
SAMEK, Wojciech, Grégoire MONTAVON, Andrea VEDALDI, Lars K. HANSEN y Klaus-Robert MÜLLER (2019): *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Cham: Springer.
SEARLE, John R. (1980): «Minds, Brains, and Programs», *Behavioural and Brain Sciences*, 3(3), 417-457.
– (2002): «Can Computers Think?», en D. Chalmers, D. (ed.) *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: Oxford University Press, 669-675. doi: 10.1037/007745.
SOURDIN, Tania (2018): «Judge v Robot? Artificial Intelligence and Judicial Decision-Making», *UNSW Law Journal*, 41(4), 1114-1133.
STARR, Sonja B. (2014): «Evidence-based sentencing and the scientific rationalization of discrimination», *Stanford Law Review*, 66(4), 803-872.
SURDEN, Harry (2014): «Machine Learning and Law», *Washington Law Review*, 89(1), 87-116.
TARUFFO, Michele (1998): «Judicial Decisions and Artificial Intelligence», *Artificial Intelligence and Law*, 6, 311-324. doi: 10.1007/978-94-015-9010-5_7.
TOULMIN, Stephen E. (2003): *The Uses of Argument*. Updated Ed, *The Uses of Argument*, Updated ed. Cambridge: Cambridge University Press.
WARREN, Roger K. (2009): «Evidence-Based Sentencing: The Application of Principles of Evidence-Based Practice to State Sentencing Practice and Policy», *University of San Francisco Law Review*, 43, 585-634.
WASHINGTON, Anne L. (2019): «How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate», *The Colorado Technology Law Journal*, 17(1), 1-37.
WISSER, Leah (2019): «Pandora's Algorithmic Black Box: The Challenges of Using Algorithmic Risk Assessments in Sentencing», *American Criminal Law Review*, 56, 1811-1832.
WOLFF, Michael A. (2008): «Evidence-based judicial discretion: promoting public safety through state sentencing reform», *New York University Law Review*, 83(5), 1389-1419
ZEDNIK, Carlos (2019): «Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence», *Philosophy & Technology* (2019). https://doi.org/10.1007/s13347-019-00382-7
ZENG, Jiaming, Berk USTUN y Cynthia RUDIN (2017): «Interpretable classification models for recidivism prediction», *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 180(3), 689–722. doi: 10.1111/rssa.12227.

Fecha de recepción: 26 de febrero de 2021.
Fecha de aceptación: 14 de mayo de 2021.