¿PUEDE LA INTELIGENCIA ARTIFICIAL SER JUSTA? ATENCIÓN ESPECÍFICA AL ÁMBITO DE LAS RELACIONES PRIVADAS

CAN ARTIFICIAL INTELLIGENCE BE JUST? A FOCUS ON THE DOMAIN OF PRIVATE RELATIONS

Rev. Boliv. de Derecho Nº 42, julio 2026, ISSN: 2070-8157, pp. 638-671

Luca MORATAL ROMÉU

ARTÍCULO RECIBIDO: 20 de junio de 2026
ARTÍCULO APROBADO: 30 de junio de 2026

RESUMEN: El presente trabajo aborda la pregunta de si es posible esperar decisiones justas de la inteligencia artificial (IA) en el ámbito de las relaciones privadas (contratación de personal, selección de proveedores, solicitudes de crédito, etc.). Como resultado del análisis, se sostiene que la IA presenta limitaciones estructurales que la colocan en una posición de desventaja frente al juicio humano, aludiéndose singularmente a su impermeabilidad a la información tácita, subjetiva y contextual cuyo procesamiento es connatural a la inteligencia humana. Sin embargo, se reconoce el valor de la IA como herramienta auxiliar, capaz de ampliar nuestras capacidades decisorias si es usada con cautela y conciencia de sus carencias, y bajo supervisión informada.

PALABRAS CLAVE: Inteligencia artificial; justicia; relaciones privadas; toma de decisiones; información; función empresarial.

ABSTRACT: *This paper addresses the question of whether it is possible to expect just decisions from artificial intelligence (AI) in the field of private relations (recruitment, supplier selection, credit applications, etc.). As a result of the analysis, it is argued that AI presents structural limitations that place it at a disadvantage compared to human judgment, with particular emphasis on its impermeability to tacit, subjective, and contextual information—elements that human intelligence processes naturally. Nevertheless, the value of AI as an auxiliary tool is acknowledged, especially its potential to enhance human decision-making capacities when used cautiously, with awareness of its shortcomings, and under informed supervision.*

KEY WORDS: *Artificial intelligence; justice; private relations; decision-making; information; entrepreneurship.*

SUMARIO: I. LA INTELIGENCIA ARTIFICIAL EN CLAVE FILOSÓFICA: MARCO Y COORDENADAS PARA SU ESTUDIO.- 1. ¿Problemas *completamente* nuevos?- 2. Dificultades, limitaciones y prejuicios.- 3. ¿Dónde nos encontramos? Lo que *sí* sabemos.- 4. La gran incógnita: el tercer estadio.- II. ALGUNAS ACLARACIONES NECESARIAS EN TORNO AL PROBLEMA DE LA JUSTICIA EN RELACIÓN CON LA INTELIGENCIA ARTIFICIAL.- III. DECISIONES JUSTAS EN RELACIONES PRIVADAS.- IV. CONCLUSIONES.

"As machine-learning systems grow not just increasingly pervasive but increasingly powerful, we will find ourselves more and more often in the position of the 'sorcerer's apprentice': we conjure a force, autonomous but totally compliant, give it a set of instructions, then scramble like mad to stop it once we realize our instructions are imprecise or incomplete—lest we get, in some clever, horrible way, precisely what we asked for".

(Brian Christian, *The alignment problem*)[1]

I. LA INTELIGENCIA ARTIFICIAL EN CLAVE FILOSÓFICA: MARCO Y COORDENADAS PARA SU ESTUDIO.

En este primer apartado propongo una aproximación general al fenómeno de la inteligencia artificial (IA) en clave filosófica, como composición de lugar necesaria para los problemas que el presente trabajo aborda. A este objeto, y con plena conciencia de la imperatividad de un importante esfuerzo sintético, reputo merecedoras de atención cuatro tareas: en primer lugar, cuestionar si, en lo que respecta a la IA, nos encontramos realmente ante desafíos inéditos, y hasta qué punto; a continuación, examinar algunas de las principales dificultades, limitaciones y prejuicios que condicionan su estudio; en tercer lugar, ubicar nuestra situación actual en el desarrollo histórico de la IA, y, por último, explorar la gran incógnita que representa su eventual evolución hacia formas de inteligencia superiores a la humana.

1. ¿Problemas *completamente* nuevos?

Si hay algo en lo que estuvieron de acuerdo Platón y Aristóteles es en que la filosofía nace del asombro (*θαυμάζειν*)[2]. Es, por tanto, más que natural que

1 Christian, B.: *The Alignment Problem. Machine Learning and Human Values*, W. W. Norton & Co., New York, 2020, pp. 12-13.

2 "[…] ese estado afectivo, el asombro (*thaumázein*), es muy propio del filósofo. En efecto, el origen de la filosofía no es otro que éste […]" (Platón: *Teeteto* [introducción, traducción y notas de Boeri, M.], Losada, Buenos Aires, 2006, 155d, p. 101); "[…] los hombres —ahora y desde el principio— comenzaron a filosofar al quedarse maravillados ante algo […]" (Aristóteles: *Metafísica* [introducción, traducción y notas de Calvo Martínez, T.], Gredos, Madrid, 1994, A, 2, 982b, 12 y ss., p. 76).

• Luca Moratal Roméu
Profesor de Filosofía del Derecho en la Universidad de Córdoba. Correo electrónico: lmoratal@uco.es.

la IA, este prodigioso fenómeno que en los últimos años ha irrumpido de lleno en nuestras vidas, se preste tanto a la reflexión filosófica; máxime cuando sus posibilidades —muchas todavía por explorar, e incluso por *imaginar*— nos asombran tanto como nos inquietan.

Pero el asombro y la inquietud, que no sólo son legítimos, sino que pueden dar lugar a consideraciones muy valiosas ante una realidad que indudablemente las requiere, pueden también llegar a cegarnos. Esto es lo que sucede cuando asumimos que los problemas que la IA plantea son completamente nuevos; o, peor aún, que exigen soluciones completamente nuevas. Lejos de ello, lo cierto es que, siendo problemas nuevos en su *species* (necedad sería negarlo), pertenecen sin embargo a un *genus* ya bastante añejo de problemas: el de los causados por la asimetría entre el desarrollo técnico de una civilización y lo que podríamos llamar —en términos, lo sé, un tanto etéreos, pero ello como necesaria consecuencia de su insoslayable amplitud— su "altura humana" (entendiendo por tal su madurez moral, espiritual, ética, filosófica, política, jurídica, cultural…).

José Ortega y Gasset lo diagnosticó magistralmente hace casi un siglo en *La rebelión de las masas*. Históricamente, la forma habitual de esta asimetría lo fue el déficit de desarrollo técnico frente al "humano"; aunque sólo sea por la lentitud con que, hasta tiempos muy recientes, ha discurrido el primero[3]. En nuestra época, en cambio, el panorama se ha revertido, y la asimetría viene ahora dada por un desarrollo humano general que, con independencia de la valoración que se haga del mismo[4], se ha quedado atrás respecto de un progreso

3 No hay más que pensar en Grecia: hasta el día de hoy, Occidente vive de sus intuiciones científicas y sus descubrimientos geométricos, lógicos y matemáticos; la manera griega de entender y, sobre todo, de ordenar el mundo, es en gran medida la nuestra; religiones tan insobornables como la judía y la cristiana no pudieron sustraerse a la influencia helénica; Whitehead escribió que "la caracterización general más segura de la tradición filosófica europea es que ésta consiste en una serie de notas al pie a Platón" (*the safest general characterization of the European philosophical tradition is that it consists of a series of footnotes to Plato*; Whitehead, A. N.: *Process and Reality [Corrected edition. Edited by Griffin, D. R., and Sherburne, D. W.]*, The Free Press, New York, [1929] 1978, p. 39), y tal vez esto pueda considerarse exagerado, pero ciertamente no lo sería caracterizar dicha tradición filosófica europea como "una serie de notas al pie a la filosofía griega"; el griego fue la lengua franca del Mediterráneo durante más de medio milenio. Y, sin embargo, nada de esto vino acompañado de una fecundidad técnica especialmente notable.

4 El tradicionalista lo juzga regresivo (célebre es el discurso pronunciado por Donoso Cortés en el Congreso de los Diputados el 4 de enero de 1849: "El fundamento, señores, de todos vuestros errores […] consiste en no saber cuál es la dirección de la civilización y del mundo. Vosotros creéis que la civilización y el mundo van, cuando la civilización y el mundo vuelven"; Donoso Cortés, J.: "Discurso pronunciado en el Congreso el 4 de enero 1849", en Donoso Cortés, J.: *Obras de don Juan Donoso Cortés, marqués de Valdegamas*, Sociedad Editorial de San Francisco de Sales, Madrid, 1892, p. 87); el relativista renuncia a juzgarlo o excluye la posibilidad de juzgarlo, lo que equivale a descartar su comparabilidad en el tiempo y en el espacio y, por consiguiente y a efectos prácticos, a afirmar su constancia en la ecuación; el "progresista" lo ve avanzar en la historia. Pero incluso este último debe reconocer que, por mucho que progresen la inteligencia, la cultura o la moral humanas, la técnica lo hace mucho más rápido; y, de hecho, es ésta una preocupación recurrente de autores más o menos generalmente considerados "progresistas" (un ejemplo nos lo proporciona Galimberti [*L'usura della terra*, AlboVersorio, Senago, 2014], al afirmar que el control sobre el desarrollo de la técnica "parece no estar ya en manos del hombre" [p. 39], toda vez que "hoy la técnica no acontece como *consecuencia de las acciones humanas*, sino como *resultado cumulativo de sus propios procedimientos*, donde los efectos se suman de tal manera que los productos finales dejan de ser reconducibles a los agentes iniciales" [p. 17]).

técnico cada vez más acelerado. Los productos de la ciencia, de la tecnología, de la industria, de la computación, etc., crecen exponencialmente en complejidad, y a un ritmo muy superior a aquél al que pueda crecer la complejidad —o la capacidad de procesamiento de complejidades— de nuestra mente; de nuestras categorías éticas, morales o jurídicas; de nuestras capacidades epistemológicas o de interpretación filosófica, etc. En otras palabras, a medida que nuestro ingenio técnico nos brinda criaturas más y más potentes, sofisticadas y admirables, nos encontramos mayores dificultades para gobernarlas, para cabalgarlas y aun para saber adónde queremos dirigirlas. O, expresándolo de manera todavía más simple: la técnica parece habérsenos ido de las manos.

[…] ahora es el hombre quien fracasa por no poder seguir emparejado con el progreso de su misma civilización. Da grima oír hablar sobre los temas más elementales del día a las personas relativamente más cultas. Parecen toscos labriegos que con dedos gruesos y torpes quieren coger una aguja que está sobre una mesa. Se manejan, por ejemplo, los temas políticos y sociales con el instrumental de conceptos romos que sirvieron hace doscientos años para afrontar situaciones de hecho doscientas veces menos sutiles.

Civilización avanzada es una y misma cosa con problemas arduos. De aquí que cuanto mayor sea el progreso, más en peligro está. La vida es cada vez mejor; pero, bien entendido, cada vez más complicada. Claro es que al complicarse los problemas se van perfeccionando también los medios para resolverlos. Pero es menester que cada nueva generación se haga dueña de esos medios adelantados[5].

Con la IA, pues, ocurre lo mismo que con tantos otros frutos de la técnica: pueden ser medios privilegiados para fines nobilísimos; pero, para que así sea, y para que no se conviertan, por el contrario, en instrumentos de iniquidad, debemos hacernos dueños de ellos. Espero, con estas reflexiones, poder contribuir a esta misión que nos incumbe a todos.

2. Dificultades, limitaciones y prejuicios.

Ante un nuevo problema (o fuente de los mismos), una nueva realidad o un nuevo fenómeno —aparte de plantearnos, como hemos hecho, que tal vez no sean tan "nuevos" como parecen—, es siempre buena estrategia tomar nota de las dificultades y limitaciones a las que nos enfrentamos a la hora de analizarlos, así como de los prejuicios que pueden condicionar nuestro análisis.

5 Ortega y Gasset, J.: *La rebelión de las masas* (1930), en Ortega y Gasset, J.: *Obras completas. Tomo IV (1929-1933)*, 6ª ed., Revista de Occidente, Madrid, (1947) 1966, p. 203.

En el caso de la IA, hay una limitación clara, y es que la mayor parte de nosotros solamente hemos interactuado con sistemas, modelos y aplicaciones[6] (ejemplos típicos son *ChatGPT* o *Gemini*) que, por prodigiosos que resulten, no dejan de ser básicos en comparación con otros mucho más perfeccionados que ya están a la orden del día en industrias punteras. Entre estos últimos cabría mencionar, en el ámbito de la medicina, *IBM Watson for Genomics*, que combina grandes volúmenes de literatura médica, datos genéticos y estudios clínicos para proponer terapias específicas y tratamientos personalizados para pacientes con cáncer; en el de la automoción autónoma, *Tesla Full Self-Driving, Waymo* o *Cruise*, capaces de procesar e integrar una infinidad de datos de toda índole en tiempo real para imitar la conducción humana; en el de la investigación científica, *Deepmind AlphaFold*, que, coordinando enormes cantidades de información, teoría evolutiva y simulaciones moleculares a través del aprendizaje constante, predice la estructura tridimensional de proteínas a partir de su secuencia de aminoácidos. Son sólo algunos ejemplos de sistemas diseñados para tomar decisiones complejas, profundamente técnicas y de alto riesgo y valor añadido, lo que implica una arquitectura altamente especializada cuyos contornos, como es natural, escapan al escrutinio del "usuario de a pie" de IA.

En defensa de este último, no obstante, pueden alzarse dos alegaciones. La primera es que los sistemas más sofisticados de IA presentan tal complejidad que, a menudo, tampoco sus usuarios, e incluso sus promotores, son capaces de hacerse una idea solvente de tales contornos (aunque, sin duda, se encuentran en una situación mucho más favorable para llegar a adquirir semejante comprensión). La segunda es que, pese a sus diferencias, todos los sistemas de IA revisten una serie de características comunes cuya observación en cualquiera de ellos (o, por lo menos, en muchos de ellos, también de los más básicos o accesibles) permite que incluso el lego en la IA más desarrollada, especializada o perfeccionada se forme una cierta representación del funcionamiento, las posibilidades y los riesgos de la IA en general o globalmente considerada. Así, ambos niveles de IA —el accesible y el más avanzado— comparten los mismos fundamentos tecnológicos, como el aprendizaje autónomo a partir de grandes volúmenes de datos, las llamadas redes neuronales (arquitecturas computacionales complejas que imitan ciertos aspectos del procesamiento humano) o el entrenamiento en datos masivos; uno y otro requieren un uso supervisado o controlado, validación por humanos expertos y contexto claro para proporcionar resultados útiles y seguros, y los riesgos —sesgo, dependencia, pérdida de tensión intelectual y capacidad crítica, mal uso intencional, etc., etc.— se hacen rápidamente evidentes en todos los planos. No quiero con esto sugerir que el usuario de *ChatGPT* esté tan autorizado para reflexionar sobre la IA como el experto en los sistemas más especializados, e insisto en que ésta

6 En adelante, para mayor simplicidad, hablaré de "sistemas de IA", entendiendo por tales cualquier tipo de herramienta, aplicación, programa, etc., cuyo funcionamiento integre el empleo de IA.

constituye para el primero una limitación incuestionable; pero no es una limitación que deba abocar al 99,9 % de usuarios de IA al silencio, sino más bien una que debe incitarnos permanentemente a enriquecer nuestros conocimientos con el estudio, en la medida de nuestras posibilidades, de aquellos usos de la IA con los que estamos menos familiarizados.

Muy relacionada con esta limitación se muestra una dificultad, consistente en la opacidad interna que es inherente a los sistemas de IA, precisamente por razón de su complejidad. Decíamos hace un momento que a menudo —y quizá no sería disparatado omitir la locución adverbial de frecuencia— ni siquiera los impulsores de los más desarrollados sistemas de IA conocen realmente los entresijos de su funcionamiento; de un modo que recuerda al jardinero que sabe qué semilla ha plantado, y cómo, y con qué regularidad y cantidad de agua ha de regarla, para que crezca determinada especie de árbol, pero que no sabría describir el proceso biológico merced al cual este árbol crecerá. A este respecto son muy elocuentes las confesiones de Fei-Fei Li, investigadora líder del proyecto ImageNet (revolucionario en aplicación de IA al reconocimiento de imágenes) y reconocida como una de las cien personas más influyentes en materia de IA por la revista *Time*[7]:

Todo lo que esta nueva generación de IA [se refiere fundamentalmente a la década de 2010] era capaz de hacer —ya fuera bueno o malo, esperado o no— se veía enrevesado por la falta de transparencia intrínseca a su diseño. La misma estructura de la red neuronal estaba entretejida de misterio: una colosal variedad de diminutas unidades de toma de decisiones, delicadamente ponderadas, carentes de sentido por separado, pero asombrosamente poderosas cuando se organizaban a gran escala, y por tanto prácticamente inmunes a la comprensión humana. Aunque podíamos hablar de ellas en un sentido teórico y distante —lo que podían hacer, los datos que necesitaban para lograrlo, el rango general de su desempeño una vez entrenadas—, lo que hacían exactamente en su interior, de una ejecución a otra, era completamente opaco. […] Conforme empecé a reconocer este nuevo panorama […] llegué a la conclusión de que las etiquetas simples ya no eran adecuadas. Incluso expresiones como "fuera de control" sonaban eufemísticas. La IA no era un fenómeno, ni una disrupción, ni un enigma, ni un privilegio. Estábamos ante una fuerza de la naturaleza[8].

Y no debemos obviar los prejuicios, que afectarán tanto a la consideración objetiva de la IA como a la valoración de sus posibilidades. Por ejemplo, la visión que se tenga, y la explicación que se aporte, de la correspondencia (o no) de los

7 Henshall, W.: "Fei-Fei Li", *Time*, 7 de septiembre de 2023, https://time.com/collection/time100-ai/6308945/fei-fei-li/ (consultado el 30 de julio de 2025).

8 Li, F.-F.: *The Worlds I See: Curiosity, Exploration, and Discovery at the Dawn of AI*, Flatiron Books, New York, 2023, pp. 284 y ss.

conceptos —y, en general, del lenguaje— manejados por un sistema de IA con las realidades a las que se refieran, habida cuenta de la incapacidad del sistema para aprehender sensorial y por tanto empíricamente estas últimas, dependerá necesariamente de la familia gnoseológica a la que se pertenezca (no serán iguales, verbigracia, las de un teoreticista que las de un descripcionista, por emplear las categorías de Gustavo Bueno[9]). Y, en el orden valorativo, nadie esperará que la IA merezca el mismo juicio ético de quien se ahorra horas y horas de trabajo gracias a ella y de quien, por su culpa, lo ha perdido. Con esta constatación no quisiera estar invitando al pesimismo. Antes bien, coincido plenamente con las siguientes palabras de Gadamer: "Los prejuicios no son necesariamente injustificados ni erróneos, ni distorsionan la verdad. Lo cierto es que, dada la historicidad de nuestra existencia, los prejuicios en el sentido literal de la palabra constituyen la orientación previa de toda nuestra capacidad de experiencia. Son anticipos de nuestra apertura al mundo, condiciones para que podamos percibir algo, para que eso que nos sale al encuentro nos diga algo"[10]; de suerte que "*los prejuicios de un individuo son, mucho más que sus juicios, la realidad histórica de su ser*"[11]. Es preciso, sí, que los prejuicios tengan la primera palabra. Pero, para no perder nuestra capacidad de búsqueda de la verdad —para no ser seres meramente históricos, sino también filosóficos—, es esencial que no tengan la última. Que no nos ofusquen. Que no nos encadenen y esclavicen. Que sean susceptibles de revocación por el tribunal supremo de la razón, aun cuando ello nos disguste. *Magis amica veritas*.

Por supuesto no he pretendido aquí, ni mucho menos, pasar revista exhaustivamente a todas las dificultades, limitaciones y prejuicios que pueden presentarse en el estudio y crítica de la IA; pero sí creo haber señalado algunos ejemplos de los más relevantes o frecuentes.

3. ¿Dónde nos encontramos? Lo que *sí* sabemos.

Tan esclarecedor como las anteriores tomas de conciencia es saber *dónde nos encontramos*: qué estadios componen la evolución del objeto de estudio y desde cuál de ellos, en nuestro momento histórico, estamos condenados a examinarlo.

Evidentemente hay en ello un ingrediente de incertidumbre, pues, por bien que conozcamos la evolución experimentada hasta el momento actual por el objeto en cuestión, no podemos más que vaticinar qué vendrá después —qué etapas y formas de desarrollo sucederán a las verificadas. Por otro lado, no podemos

9 Cf. Bueno, G.: *Teoría del cierre categorial. Volumen I: Introducción General. Siete enfoques en el estudio de la Ciencia*, Pentalfa, Oviedo, 1992, pp. 57 y ss..

10 Gadamer, H.-G.: *Verdad y método II* (traducción de Olasagasti, M.), Sígueme, Salamanca, (1965/66) 1998, p. 218.

11 Gadamer, H.-G.: *Verdad y método I* (traducción de Agud Aparicio, A., y De Agapito, R.), 8ª ed., Sígueme, Salamanca, (1960) 1999, p. 344 (cursiva en el original).

sencillamente ignorar ese vector futuro de desarrollo, no al menos si, en efecto, queremos construirnos una idea sólida de nuestra situación presente (o de la del objeto de estudio). Así pues, no nos queda más remedio que preguntarnos qué cabe razonablemente esperar, a la luz de la evolución precedente de la realidad estudiada y de otras afines; y ello —esto es clave— desde el prisma de las categorías y criterios que interesen a nuestro estudio. Por ejemplo, una línea evolutiva destinada a determinar en qué fase de desarrollo de los antibióticos nos encontramos será muy distinta en función de que adoptemos como criterio de fasificación: i) el grado de complejidad química de los mismos, ii) los niveles de resistencia patogénica a antibióticos que sean capaces de vencer, iii) las categorías de enfermedades a las que puedan aplicarse… Dependiendo del criterio adoptado, el estadio "antibióticos de primera generación" (o "de segunda", o "de tercera", etc.) tendrá un significado y un alcance distintos.

En el terreno de la IA son también muchos los criterios asumibles. No obstante, creo que el más interesante de estos criterios, si a lo que se aspira es a un panorama general del impacto de la IA sobre el devenir humano, es *la parangonabilidad de la IA con la inteligencia humana*; criterio conforme al cual podría sencillamente dividirse la evolución de la IA (como, por lo demás, comúnmente se hace) en:

1- IA de capacidad inferior a la inteligencia humana (*subhuman-level AI*).

2- IA de capacidad pareja a la inteligencia humana (*human-level AI*, *par-human AI*).

3- IA de capacidad superior a la inteligencia humana (*superhuman-level AI*).

Esta clasificación no está exenta de problemas. Uno de ellos es qué entendemos por "capacidad", o incluso por "inteligencia" (¿debemos considerar como una de sus dimensiones la "inteligencia emocional"?). Definido esto, habría que tener en cuenta que una misma IA puede ser a veces *subhuman* y a veces *par-human* —¿y otras veces es, o podrá ser, *superhuman*?—, lo que dificulta seriamente el establecimiento de límites fijos entre los tres estadios. Por otra parte, y dado que el criterio es una relación con la inteligencia humana, habría que plantearse si el punto de referencia a adoptar para valorar esta última es un individuo concreto (y, en tal caso, ¿con qué características?: ¿las de un individuo medio?) o la humanidad en su conjunto, y si, a efectos de testeo comparativo, ese punto de referencia dispondría del mismo tiempo que la IA a clasificar para la realización de una misma tarea (¿y qué tarea habría de ser ésta?). Otra complicación es que, al no (poder) saber cómo será la *superhuman-level AI* —ni si llegará—, desconocemos igualmente si será necesario distinguir en ella subfases, acerca de las cuales no podríamos más que elucubrar sin ningún tipo de fundamento.

Objeciones como éstas aconsejan, ciertamente, no llevar demasiado lejos esta clasificación. Sin embargo, tampoco podemos eludirla. La magnitud revolucionaria de la IA vendrá dada por su aptitud para imitar e incluso, llegado el caso, superar el poder de la inteligencia humana. Si esto es innegable, la consecuencia es que no podemos sustraernos a valorar su evolución desde este punto de vista —es decir, a través de una clasificación como ésta—, por mucho que podamos y debamos discrepar en torno a los extremos enunciados y muchos otros. Por lo demás, el escollo de no saber qué nos deparará el futuro subsiste también, cómo no, en este frente. No obstante, la generalidad de esta clasificación contribuye a superarlo ampliamente: en efecto, no parece aventurado reputar inconcebible un estadio evolutivo de la IA no subsumible en alguno de estos tres (sin perjuicio de situaciones a caballo entre uno y otro y otros problemas apuntados, como la relatividad de las variables).

Así las cosas, y no sin las debidas reservas, podemos hacer nuestra esta clasificación para responder al interrogante de dónde nos encontramos, o dónde se encuentra la IA respecto de nosotros: ¿por debajo?, ¿más o menos a la par?, ¿por encima? Algo parece claro, y es que no estamos en esta última fase (*superhuman-level AI*), que es —con razón— la que más preocupa, tanto por su potencial (al servicio del bien, pero también del mal) como por su imprevisibilidad (imprevisibilidad que viene impuesta por su propia definición; pues, desde el momento en que las capacidades de la IA sean superiores a las nuestras, sería vano pretender predecir su comportamiento desde nuestra inferioridad intelectiva, analítica, etc.). Pero es que probablemente ni siquiera estamos todavía en la segunda fase (*human-level AI*, *par-human AI*). Así lo entendía en marzo de 2025 Demis Hassabis, Premio Nobel de Química en 2024 (por el desarrollo de la herramienta *AlphaFold*, a la que nos hemos referido anteriormente), al pronosticar que la IA "igualará la inteligencia humana dentro de los próximos 5 a 10 años" (*will match human intelligence within the next 5–10 years*)[12].

En mi propia experiencia (limitada a modelos básicos, sí, pero creo que bastante representativa, a juzgar por el dictamen de Hassabis), la IA sólo es capaz de emular un trabajo humano de calidad bajo supervisión estricta, continuada y activa[13]. Sin ella, lo normal es que la IA haga mucho más rápido lo que muchos humanos mediocres habrían hecho mucho más lentamente; y que lo haga de manera formalmente impecable, pero con la misma mediocridad material. A poco que lo pensemos, nos damos cuenta de que lo que nos asombra de la IA no

12 Loeb, A.: "Superhuman AI is Closer Than We Think, for the Wrong Reason!", *Medium*, 30 de marzo de 2025, https://avi-loeb.medium.com/superhuman-ai-is-closer-than-we-think-for-the-wrong-reason-0fd80dcf2a83 (consultado el 30 de julio de 2025).

13 Y quizá en muchos casos ni siquiera así sea capaz de emularlo, pues habría que tener en cuenta una referencia que casi nadie baraja casi nunca (precisamente porque *casi siempre* usamos la IA para ahorrarnos trabajo y tiempo, no para contrastarnos): el coste de oportunidad, a saber, cómo un humano habría hecho el mismo trabajo (aquí volvemos a encontrarnos con el problema de *en cuánto tiempo*) sin ayuda de IA.

es su genialidad, sino sobre todo su autonomía, su celeridad y su capacidad de mimetismo con el trabajo humano. Y tan predicable es esto de la IA predictiva (programada para reconocer patrones en datos históricos y estadísticos con el fin de anticipar o pronosticar eventos futuros) como de la generativa (creadora de contenido nuevo a partir de instrucciones).

Estas impresiones se ven respaldadas por algunas cosas que *sí* sabemos —elenco que palidece frente al de las cosas que *no* sabemos; pero elenco que no deberíamos menospreciar— acerca de la IA.

La primera de las que nos interesan es cuál es la materia prima que procesa la IA a la hora de realizar cálculos, emitir valoraciones, aportar soluciones, tomar decisiones, etc. Dicha materia prima somos nosotros, o, más exactamente, lo que podríamos llamar —si se me permite la terminología marxista— nuestra "superestructura": nuestro comportamiento plasmado en algún tipo de lenguaje o registro (generalmente verbal, pero también matemático, estadístico, etc.), incluyendo las instrucciones que nosotros mismos le damos a la IA para efectuar las operaciones que le encomendamos. Así pues, es verdad que esta materia prima no somos nosotros en nuestra integridad, y ni siquiera en nuestro ser más fundamental (nuestra "infraestructura"), pero sí, a fin de cuentas, nuestra huella en este mundo, y por lo tanto algo fielmente reconducible a nuestro ser en sentido pleno. Así como la huella no es nuestro pie, pero se corresponde con sus medidas y su forma, y proporciona una idea bastante nítida de cómo es nuestro pie, la plasmación de nuestro comportamiento en registros o lenguajes legibles y analizables por los sistemas de IA (plasmación existente en volúmenes masivos, gracias a Internet) ofrece a estos últimos un cuadro altamente fiel de quiénes somos los humanos, y de cómo pensamos, funcionamos y actuamos. Podemos, pues, afirmar que la IA se nutre de nosotros. Pero también podemos afirmar que hay una parte de nosotros —la "infraestructura" humana, semilla y motor de nuestra "superestructura", así como aquella parte de esta última que, no habiendo quedado codificada, o habiendo desaparecido, no puede ser procesada por la IA— que queda fuera de su alcance. Ni una cosa ni la otra son argumentos definitivos a propósito de que nos encontremos todavía en un estadio de *subhuman-level AI* (pues bien podría ser, desde una perspectiva hilemorfista, que lo humano o aun lo suprahumano de la IA fuera, no ya la materia, sino la forma), pero ambas dicen algo a favor de esta tesis.

Otra cosa que conocemos, y que también puede resultar reveladora al objeto de ubicarnos, son algunas generalidades de la dinámica de aprendizaje de los sistemas de IA. Christian las describe en los siguientes términos:

El campo del aprendizaje de máquinas comprende tres áreas principales: En el aprendizaje no supervisado, se le da a la máquina un cúmulo de datos y […] se

le pide que les dé sentido, que encuentre patrones, regularidades, formas útiles de condensarlos, representarlos o visualizarlos. En el aprendizaje supervisado, al sistema se le proporciona una serie de ejemplos categorizados o etiquetados —como personas en libertad condicional que fueron arrestadas nuevamente y otras que no lo fueron— y se le pide que haga predicciones sobre nuevos ejemplos que aún no ha visto, o para los cuales la verdad empírica aún no se conoce. Y en el aprendizaje por refuerzo, el sistema se coloca en un entorno con recompensas y castigos […] y se le pide que descubra la mejor manera de minimizar los castigos y maximizar las recompensas[14].

De acuerdo con esto —y en la medida en que no se haya superado esto—, el protagonismo humano no está solamente en la materia de la que se nutre la IA, sino igualmente en la *forma*; esto es, en los criterios, los parámetros, los valores, las instrucciones, los objetivos… y, en definitiva, la lógica, que rigen el procesamiento artificial de aquella materia. En todo ello, la IA seguiría subordinada a la inteligencia humana.

Soy consciente de que la cuestión requeriría un examen mucho más profundo, pero, como conclusión de lo expuesto, mi apuesta es que, en el momento presente, nos encontramos todavía inmersos en el primer estadio evolutivo de la IA (*subhuman-level*); no sin destellos, ahora bien, de ese segundo estadio (*human-level*) que, según Hassabis, podría estar muy cerca.

4. La gran incógnita: el tercer estadio.

Si realmente nos encontramos todavía en un estadio de IA de nivel inferior o, a lo sumo, puntualmente parejo al de la inteligencia humana, mas donde el protagonismo lo sigue detentando esta última, una regla básica debería regir nuestra actitud ante, y nuestras relaciones con, la IA: no temerla más a ella de cuanto nos temamos a nosotros mismos. Lo peligroso, hoy por hoy, no es la IA en sí misma, o lo que pueda hacer por sí misma, sino el material que los humanos le proporcionamos y el uso que podamos hacer de ella. Esto, se dirá, no es poco motivo de alarma; pero es "malo conocido" que, según el adagio popular, es mejor que "bueno por conocer".

La gran incógnita es el tercer —¿y último?— estadio. Decíamos antes que éste es imprevisible por definición, pues toda previsión que pudiéramos hacer de él partiría necesariamente de unas facultades intelectuales (también volitivas) que, como estamos precisamente asumiendo, se verían superadas por el destinatario de tales previsiones (la IA de nivel superior a la inteligencia humana), con lo que sería un tanto ingenua la pretensión de efectuarlas. No sabríamos ni siquiera concebir

14 Christian, B.: *The Alignment Problem*, cit., p. 47.

en qué puedan consistir, cómo vayan a operar o de qué sean capaces semejantes facultades, de las cuales no tenemos experiencia alguna. Así, si representan una amenaza para la especie humana (lo que no es ni mucho menos seguro), no podríamos saber cómo combatirla. Por otro lado, tampoco existe garantía ninguna de que ese estadio tenga que llegar; ni, si tiene que llegar, de que nosotros lleguemos a presenciarlo. Que el horizonte teórico en el que nos movemos sea éste —por el simple hecho de que, si vislumbramos ya como factible a corto o medio plazo una *human-level AI*, debemos trabajar con la hipótesis de un posible estadio ulterior, que necesariamente sería el de una *superhuman-level AI*— no garantiza nada en este sentido.

En este punto es inevitable que se plantee el debate de si merece la pena consentir un desarrollo a largo plazo de la IA que pueda eventualmente conducir al tercer estadio, corriendo el riesgo —de dimensiones inciertas— de que éste, además de ser realizable —y sabiendo que, en tal caso, será incontrolable—, nos sea hostil. Pero, ¿hay alternativa? ¿Estamos en condiciones de abortar el desarrollo de la IA? ¿Quién podría hacerlo? Los Estados y algunas organizaciones supranacionales (como la Unión Europea) tienen el poder de prohibírselo al grueso de sus ciudadanos, empresas, etc.; pero los hay (pensemos en las empresas de mayor músculo financiero, mejores relaciones institucionales…) que podrán eludir y de hecho eludirán la prohibición. Lo más significativo, en cualquier caso, no es esto, sino la cuestión de quién impediría a los Estados —que entre sí, como sabemos por Hobbes y Locke, se encuentran en estado de naturaleza[15]— seguir desarrollando la IA para sus propios fines; pues es evidente que, si la IA puede ser peligrosa en manos de los particulares, los riesgos se multiplican en manos de los Estados. Quizá los más poderosos pudieran restringir el acceso a ella de los más débiles (aunque tampoco esto sería nada fácil), pero sería ilusorio pensar que las grandes potencias vayan a estar dispuestas en algún momento a renunciar a las oportunidades (represivas, propagandísticas, geopolíticas…), presentes y futuras, que les ofrece la IA. Aunque sólo fuera por esta razón, el avance de la IA es imparable, y también el presente debate carece de sentido.

II. ALGUNAS ACLARACIONES NECESARIAS EN TORNO AL PROBLEMA DE LA JUSTICIA EN RELACIÓN CON LA INTELIGENCIA ARTIFICIAL.

No menos importante que manejar una noción del primero de los términos de nuestro binomio (la IA) es poseer una idea del segundo (la justicia). En este sentido, antes de entrar a fondo en materia es preciso resolver algunas dificultades preliminares que afectan al mismo concepto de justicia y a su proyección en este terreno.

15 Cf. *Leviathan*, Chap. XXX *in fine* (Hobbes, T.: *Leviathan*, Penguin, London, [1651] 2017, p. 292), y *Second Treatise of Government*, Chap. XII, 145 (Locke, J.: *Second Treatise of Government and A Letter Concerning Toleration*, OUP, Oxford, [1689] 2016, pp. 73-74).

La literatura que se viene ocupando de esta cuestión tiende a abordarla en dos ámbitos claramente diferenciados: el público y el privado. Por una parte, se habla de la (in)capacidad de la IA para tomar decisiones justas con motivo de un pleito, a la hora de determinar si un preso debe ser puesto en libertad condicional, en la resolución de un expediente administrativo (por ejemplo, una solicitud de subsidio por desempleo), etc. Por otra parte, se habla de la (in)capacidad de la IA para tomar decisiones justas en materia de contratación de personal, ascensos, despidos, valoración de solicitudes de hipotecas, selección de proveedores o socios u otras decisiones empresariales, etc. En el primer caso se evalúa la aptitud de la IA para sustituir a un agente de los poderes públicos (el juez, el funcionario al servicio de la administración), mientras que en el segundo se examina su aptitud para hacer las veces de un particular en el seno de una relación privada. Ambas esferas revisten gran interés. Por limitaciones de espacio, en el presente trabajo trataremos específicamente la segunda, posponiendo la primera a estudios sucesivos.

Pues bien, como se prevenía, es preciso disipar algunos malentendidos a los que esta división puede dar lugar. Los equívocos se presentan fundamentalmente en lo relativo a la posibilidad de calificar como "justa" o "injusta" una decisión en el ámbito privado, del tipo de las que se han puesto como ejemplo. No quiere con esto enclaustrarse el concepto de justicia en los dominios del derecho público. Si la justicia, como reza su definición clásica, consiste en "dar a cada uno lo suyo" (*suum cuique tribuere*), se hace patente que no podrá desterrarse de las relaciones privadas, en las cuales constantemente las partes pueden y deben dar "lo suyo" a una contraparte: será justo que el vendedor entregue la cosa vendida al comprador en las condiciones pactadas, que el vecino pague su cuota de comunidad, que el padre o la madre paguen la educación de sus hijos… e injusto lo contrario. Ahora bien, en tales situaciones poco o nada tiene que decir —que *decidir*— la IA. En ellas, la justicia se concreta en el cumplimiento de alguna obligación establecida por la naturaleza (las obligaciones de alimentos entre parientes, quizá la asistencia a quien puede ser socorrido sin riesgo para el que la presta) o por el consentimiento válido (las obligaciones nacidas de un contrato). Cuestión distinta es que se trate de una obligación indeterminada o disputada, en cuyo caso sí puede plantearse si la IA sería capaz de resolver con justicia la indeterminación o la controversia (como lo haría un juez o un árbitro); pero obsérvese que, en semejante escenario, ya estaríamos de nuevo en el ámbito de lo público.

En efecto, no es esa dimensión de la justicia la que se suele indagar cuando se habla de la capacidad de la IA para adoptar decisiones justas en el ámbito privado. Más bien la referencia suele serlo a decisiones del tipo de las mencionadas a título ejemplificativo: la decisión de contratar a un candidato o a otro, de ascender a uno u otro empleado, etc., etc. Y aquí es donde sí nos asalta con toda su fuerza

la pregunta de hasta qué punto tiene sentido la categoría de la justicia aplicada a tales decisiones. ¿Puede sostenerse que haya, en los contextos en los que tienen que tomarse, un "lo suyo" de cada uno?

La opinión más extendida es que *sí*. Si preguntáramos a una muestra representativa de la población española o de cualquiera de nuestro entorno, probablemente una mayoría abrumadora de los encuestados se pronunciaría en el sentido de que, para cada una de estas necesidades decisionales, existe una decisión objetivamente más justa; o, lo que es lo mismo, un candidato objetivamente más cualificado o solvente para ser el elegido (para cubrir una vacante, para prestar un servicio, para recibir dinero a crédito…); de suerte que, mereciéndolo objetivamente, sería injusto que no lo obtuviera, en tanto que no se le estaría dando "lo suyo". Los resultados, presumo, no serían muy distintos si los interrogados fueran exclusivamente académicos o expertos en disciplinas tales como economía, sociología, gestión de recursos humanos, filosofía, derecho del trabajo o cualesquiera otras que pudieran mostrarse relacionadas con la cuestión. Es más, una posición semejante ha vertebrado la teoría económica clásica, de Adam Smith a Karl Marx, y sigue siendo preponderante a día de hoy: se trata de la concepción objetiva del valor. En su versión marxista, por ejemplo, el valor de un producto o mercancía viene determinado por el número de horas de trabajo "incorporadas" al mismo, lo que permite evaluar objetivamente si el precio (también expresado en horas de trabajo) pagado por él —o el que recibirá el trabajador a cambio de su trabajo para producirlo— es, o no, justo. De la misma manera, muchos consideran —o considerarían, si se les estimulara dialécticamente— que el mérito de los candidatos a verse beneficiados por determinada decisión depende de factores objetivos u objetivables (como, por qué no, el mismo número de horas de trabajo, de antigüedad en actividades relevantes para la vacante a cubrir). La objetivación de estos criterios puede resultar más o menos complicada, pero, una vez efectuada, tales criterios permitirían definir en cada caso cuál es la decisión objetivamente más justa.

Semejante perspectiva, ahora bien, presenta tantas y tan insalvables taras como la concepción objetiva del valor en teoría económica. Si lo que se pretende, pongamos, es el cálculo del mérito en base a la antigüedad, pueden esgrimirse objeciones similares a las alzadas por Ludwig von Mises contra la teoría marxista del valor-trabajo:

[…] la hora de trabajo no es una cantidad uniforme y homogénea. En efecto, no existe un "factor trabajo", sino innumerables tipos, categorías y clases distintas de trabajo […]. No se trata tan sólo de que la eficiencia varíe enormemente de unos trabajadores a otros, e incluso para cada trabajador según que el momento y las circunstancias y condiciones en que se desarrolla su trabajo sean más o menos

favorables, sino que, además, las clases de servicios que proporciona el factor trabajo son tan variadas y están modificándose de forma tan continuada que, de hecho, constituyen tipos absolutamente heterogéneos de servicios […][16].

En efecto, la pretensión de objetivar el mérito en la antigüedad —análogamente a la de objetivar el valor en las horas de trabajo— ignora necesariamente la esencial heterogeneidad de dicha antigüedad: ¿valen lo mismo los dos años de antigüedad del trabajador más productivo y del menos productivo?, ¿quién nos garantiza que, en ese tiempo, ambos trabajadores han adquirido y cultivado las mismas habilidades?, ¿pueden compararse esas antigüedades a pesar de haberse acumulado en entornos (empresas, lugares, sectores…) potencialmente distintos? Y, si de lo que se trata —verbigracia— es de cubrir una vacante cuya naturaleza no se corresponde exactamente con la de las actividades desarrolladas hasta el momento, ¿qué coeficiente correctivo habría que aplicar a la antigüedad alegada para obtener un resultado "justo"?

Adviértase, en este punto, que una concepción objetiva del mérito sale incluso peor parada que la teoría económica del valor-trabajo. Marx intentó defender la segunda arguyendo aquello de que el tiempo de trabajo a tener en cuenta no era el de un trabajador u otro, individualmente considerados (lo cual da lugar a resultados manifiestamente absurdos; por ejemplo, supondría equiparar el valor de un cuadro pintado por Antonio López y el de otro pintado por el autor de estas líneas, simplemente porque ambos hubiesen tardado lo mismo en completarlo), sino el tiempo de trabajo "socialmente necesario" para la producción de la mercancía[17]. Una concepción objetiva del mérito ni siquiera puede acogerse a semejante efugio, puesto que no se trata aquí de valorar un bien (que lo mismo podría haber sido producido por una persona que por otra) sino la cualificación o idoneidad de una persona concreta (que, como es obvio, únicamente tiene sentido considerar respecto de esa misma persona). Desde un punto de vista marxista puede ser aceptable afirmar que el valor de un par de zapatos no debe determinarse teniendo en cuenta las horas de trabajo empleadas para confeccionarlos por el trabajador más holgazán ni por el más productivo, sino por un trabajador medio en circunstancias normales; pero no hay manera razonable de extender semejante ajuste a la valoración del mérito personal. Podemos, sí,

16 Huerta de Soto, J.: *Socialismo, cálculo económico y función empresarial*, 3ª ed., Unión Editorial, Madrid, (1992) 2005, p. 205.

17 Cf. Marx, K.: *El capital. Crítica de la economía política. Tomo primero. Libro 1. Proceso de producción del capital*, Progreso, Moscú, (1867) 1990, pp. 48 y ss. Tampoco con esta precisión se sostiene su teoría. Como explica Huerta de Soto, "tal reducción de las horas de los diferentes tipos o clases de trabajo a las horas de trabajo [socialmente necesario] sólo es posible que se efectúe cuando existe un proceso de mercado en el cual unas y otras son intercambiadas a un precio determinado por los diferentes agentes económicos. En ausencia de este proceso de mercado, cualquier juicio comparativo sobre distintos tipos de trabajo habrá de ser arbitrario, y ello implicará forzosamente la desaparición del cálculo económico racional. El problema consiste, pues, en que no es posible reducir los diferentes tipos de trabajo a un común denominador sin que previamente exista un proceso de mercado" (*Socialismo*, cit., pp. 205-206).

asumir *iuris et de iure* que un día de antigüedad vale, siempre y en todo caso, X, haciendo abstracción del efectivo aprovechamiento que cada trabajador haya hecho del mismo; pero con ello no estaríamos evaluando la aptitud real de cada candidato. Sencillamente nos estaríamos engañando.

A veces se piensa que, dada la falibilidad de algunos criterios de objetivación (como hemos visto respecto de la antigüedad), la solución pasa por añadir más criterios a la ecuación (en el caso de un proceso de selección, titulaciones académicas, notas medias, dominio de idiomas, desempeño en pruebas *ad hoc*, buenas referencias…). Lo cierto, sin embargo, es que todos ellos presentan las mismas limitaciones, derivadas de las mismas heterogeneidad, inmensurabilidad y subjetividad de los aspectos a los que se refieren. La inclusión de variables adicionales, de hecho, no hace más que dificultar la objetivación, pues con ella surge el problema, igualmente irresoluble, de qué peso dar a cada una de ellas.

Habrá quien, concediendo estas aporías, replique que de ellas no se sigue una esencial falta de objetividad del mérito personal, sino únicamente nuestra incapacidad para aprehenderla. Dicho mérito sería realmente objetivo, pero la sutileza de sus matices lo haría inaccesible a la inteligencia humana. A nuestros efectos, empero, semejante platonismo no cambia nada, pues, si de lo que se trata es de determinar si a un candidato (por mantener este ejemplo, entre muchos otros posibles) se le ha hecho justicia, esto es, si se le ha dado "lo suyo" o "lo que merece", ¿cómo aducir que no se le ha dado —que no se le ha hecho justicia—, si "lo suyo" o "lo que merece" es universalmente incognoscible (para el que lo aduzca, para el decisor, para el encargado de resolver la controversia, etc.)? *Ad impossibilia nemo tenetur.*

Con todo, el problema de fondo no es la incognoscibilidad del mérito humano. El problema de fondo es que este mérito —que, objetivado, quizá, y sólo quizá, pues habría que superar muchas otras objeciones, nos permitiría hablar de "decisiones justas" e "injustas" en el ámbito privado— es, en el más sustancial de sus estratos, tan subjetivo como el valor económico de los bienes y servicios. Un prestamista, por las más variadas e íntimas razones, puede elegir prestar dinero al menos solvente de quienes se lo piden. Un empleador puede intuir en el candidato "objetivamente" menos cualificado un talento, actitud o sintonía que, sin saber explicar por qué, le haga preferirlo a todos los demás. Una mujer puede romperle el corazón al más abnegado de sus pretendientes, para casarse con el más indolente y desidioso. Personalmente, puedo entender que decisiones como éstas sean calificadas de "inadecuadas", "irracionales", "absurdas" y muchas cosas peores. Pero, ¿puede realmente decirse que sean "injustas"? En caso afirmativo, deberíamos ser coherentes y admitir la intervención judicial en decisiones tan personales como la elección de cónyuge, única manera de "hacer justicia" a los pretendientes "objetivamente más cualificados".

En campos como la administración de empresas, la psicología o la gestión de recursos humanos, la literatura académica lleva décadas diseñando fórmulas, modelos, matrices, baremos y todo tipo de tentativas de dotar de objetividad a la valoración de méritos en procesos de selección[18]; de igual manera que las entidades financieras disponen de los más sofisticados sistemas para determinar a quiénes conceder hipotecas y préstamos, e *via dicendo*. Con esta disertación no pretendo negar la utilidad de estos artificios para ayudar a tomar decisiones, pero sí oponerme a todo conato de absolutización filosófica, científica o jurídica de los mismos. Su utilidad, en este sentido, dependerá de la percepción subjetiva que de la misma tenga el encargado de tomar la decisión en cuestión: ni más, ni menos. Y si este decisor, en un momento dado y por cualquier motivo, se convence de que otro criterio debería prevalecer sobre el veredicto de tales artificios, ello podrá merecer muy distintos adjetivos, pero ninguno de ellos debería ser "justo" o "injusto", al menos en un sentido primario o propio.

El decisor, naturalmente, podrá e*quivocars*e, entendiendo por tal arrepentirse posteriormente de la decisión tomada. Pero esto no hace la decisión "injusta", sino sencillamente desacertada. Y, en cualquier caso, ¿quién mejor que él, que a fin de cuentas será el principal beneficiario de su acierto, y el principal damnificado por su desacierto, para tomarla? Querer reemplazarlo en esta tarea por una fórmula supuestamente "objetiva" es una manifestación más de la "fatal arrogancia" de los planificadores sociales, tan denunciada por Friedrich Hayek sobre la estela de Adam Smith[19]. Y es que, por bien construida que esté esa fórmula, siempre carecerá de dos elementos necesariamente determinantes de la superioridad del juicio privado del decisor: la información que éste posee por el mero hecho de hallarse en la posición que requiere la toma de una decisión (información de la que, como posteriormente veremos, muchas veces ni siquiera es plenamente consciente, y que a menudo es tácita, inarticulable e intransmisible) y su interés personal directo en tomar la mejor decisión posible (aquélla que maximice los beneficios a los que aspira y/o minimice los riesgos y perjuicios que anhela ahuyentar).

Algunas voces serán menos exigentes —menos "platónicas"— y reconocerán la improcedencia de calificar de "injusta" una decisión personal *siempre y cuando no sea discriminatoria*; es decir, siempre y cuando no se funde en motivos arbitrarios, tales como la raza, el sexo, la religión… Pero, ¿qué motivos son arbitrarios y cuáles no? Toda decisión, a fin de cuentas, implica una discriminación. La primera acepción que el Diccionario de la RAE (2014) atribuye al verbo "discriminar" es

18 *Vid*., e.g., la matriz de habilidades (*skills matrix*) de Dessler (Dessler, G.: *Human Resource Management*, 16ª ed., Pearson, New York, [2015] 2020, p. 124).

19 Si bien el título del último libro de Hayek, *The Fatal Conceit: The Errors of Socialism* (1988), es un inequívoco homenaje a Adam Smith, éste nunca empleó la frase exacta. En *The Theory of Moral Sentiments* (H. G. Bohn, London, [1759] 1853, pp. 342-343), Smith habla de la "arrogancia" (*conceit*) del gobernante "a menudo tan enamorado de la supuesta belleza de su propio plan de gobierno ideal que no puede tolerar la mínima desviación de cualquier parte del mismo".

"seleccionar excluyendo" (única manera, a poco que se medite, de "seleccionar"). El término proviene del latín *discrimen* (división, separación), a su vez proveniente del verbo *discerno* (discernir, dividir, separar, diferenciar). El banco que concede una hipoteca al solicitante que presenta un avalista y en cambio la deniega al que no lo aporta está discriminando (*i.e.*, discerniendo), al igual que el empresario que contrata al candidato que cree mejor capacitado para cubrir una vacante o la mujer que se casa con quien prevé que será el mejor padre para sus hijos. Puesto que toda decisión es discriminatoria, ¿qué criterios de discriminación son admisibles, y cuáles no? Definirlos objetivamente nos retrotrae a todas las dificultades anteriormente expuestas. Incluso aquellos criterios que tendemos a reputar moralmente aberrantes como determinantes de una discriminación (la raza, el sexo, la religión…) pueden resultar comprensibles en algunos contextos: ¿Es "injusto" el anunciante de lencería femenina que sólo contrata mujeres como modelos para aquélla? ¿Es "injusto" el hombre que ni siquiera contempla mantener una relación sentimental con una mujer de otra raza, o con otro hombre, o que nunca se casaría con una mujer de religión distinta a la suya? Insisto en que no se trata de emitir un juicio moral, sino específicamente de valorar si es éste, o más bien no, el terreno ideal para hacer uso de la categoría "justicia". E insisto en que, si la respuesta es afirmativa, quizá la coherencia demandaría que nuestro sistema jurídico permitiera la fiscalización de semejantes decisiones.

Estoy lejos de negar que determinadas discriminaciones, también en el ámbito privado, resultan éticamente deleznables; pero no siempre lo que nos repugna ética o moralmente es injusto. Ello no significa que la sociedad no pueda o no deba reaccionar ante este tipo de comportamientos y situaciones: lo normal es que quienes discriminan sin razón éticamente admisible, como puro acto de hostilidad contra un grupo humano, se hagan acreedores del repudio social. No es un buen negocio, en términos reputacionales. Pero tampoco lo es en un plano estrictamente económico. Discriminar de esta manera no sólo es reprobable desde muchos puntos de vista, sino también autolesivo para quien lo hace: por ejemplo, el empresario que rechaza al candidato más cualificado simplemente por su color de piel está renunciando, irracionalmente, a maximizar los beneficios de su decisión[20]. Sea como fuere, sí considero que, siendo ésta una esfera ajena a la propia de la justicia —a aquélla en la que puede hablarse, con propiedad, de decisiones y actuaciones justas e injustas—, sería pertinente que los responsables de "hacer justicia", o de administrarla, se abstuvieran de interferir en ella.

Es precisamente la irracionalidad de las discriminaciones que suelen escandalizarnos la que me hace pensar que no son tan frecuentes como a

20 Tal vez, en su escala de valores, el malsano placer de despreciar a una persona de otra raza pese más que el beneficio económico que podría haber derivado de su decisión. En tal caso, lo irracional no es tanto su decisión cuanto su escala de valores.

veces se nos quiere hacer creer: el egoísmo es el mejor antídoto para actitudes como la xenofobia, la homofobia o el sexismo. Sin embargo, no es raro que se censuren como discriminatorias decisiones y situaciones que, en realidad, no son equiparables a una discriminación sustancialmente motivada por semejantes actitudes. Un ejemplo nos lo da Schellmann:

> Las personas que están más activas en la plataforma, por ejemplo, enviando mensajes a reclutadores —algo que, en promedio, los hombres hacen con más frecuencia que las mujeres— probablemente serán recomendadas con mayor frecuencia a los reclutadores por la IA. La IA no "sabe" quién es mujer y quién es hombre, pero detecta estos patrones de comportamiento relacionados con el género [*gendered behavioral patterns*] y puede, de manera involuntaria, recomendar a más hombres que mujeres a los reclutadores basándose en esos patrones[21].

Esto no se sostiene. Si por *gendered behavioral patterns* se entiende —como efectivamente es el caso— datos conductuales en los cuales se observa un comportamiento distinto entre hombres y mujeres, prácticamente todos lo son, pues difícil será encontrar alguna manifestación de la conducta humana en la cual hombres y mujeres actúen exactamente igual, o en la que la observación de uno y otro sexo arroje idénticos resultados porcentuales. Por esta regla de tres, casi ningún *pattern* sería aceptable, pues casi todos serían discriminatorios[22]. Schellmann menciona el patrón "contactar directamente con reclutadores" (*messaging recruiters*). Sin embargo, otros patrones que la IA puede tener en cuenta (ejemplo ficticio, pero factible: nivel de luminosidad de la foto de perfil) podrían beneficiar a las candidatas.

Estas aclaraciones parecen incompatibles con mi intención declarada de tratar el problema de la inteligencia artificial y su aptitud para la justicia en el ámbito de las decisiones *privadas*; y, efectivamente, lo son, siempre que se quiera hablar de justicia en sentido estricto. No obstante, y como ocurre con tantos otros conceptos, muchas veces se maneja el de justicia en un sentido derivativo, metafórico o asimilado; en una palabra, en un sentido impropio. Hacerlo no es necesariamente ilegítimo, pues, cuando una comunidad lingüística, de manera espontánea, se decanta por estos usos derivativos, lo que pone de manifiesto es que los conceptos en cuestión son parcialmente expresivos de las realidades a las cuales son extrapolados. El peligro semántico nace de tomar la parte por el todo, pretendiendo hacer de las acepciones derivativas de un concepto su significado

21 Schellmann, H.: *The Algorithm: How AI Decides Who Gets Hired, Monitored, Promoted, and Fired and Why We Need to Fight Back Now*, Hachette Books, New York, 2024, pp. 37-38.

22 La identidad de comportamiento entre sexos sólo podría ser fruto de la casualidad, y cuanto más amplia fuera la muestra más difícil sería que éstas se produjeran.

primario o fundamental[23]. No obstante, una vez se es rectamente sabedor de cuál es este sentido primario del concepto y cuál o cuáles son los derivativos, puede darse juego a estos últimos sin riesgo de deformación. En nuestro caso, sabiendo ya que el concepto de justicia no es propiamente aplicable a las decisiones de índole privada, no habría problema en atender al hecho de que, en el lenguaje común, también a ellas se les aplica —pese a resultar más apropiados otros calificativos (decisiones acertadas y desacertadas, etc.)—; para, en base a ello, examinar si la IA sería capaz de tomar ese tipo de decisiones en el modo que la comunidad lingüística —impropia pero consistentemente— acostumbra llamar "justo". Aun cuando sólo fuera *for the sake of argument*.

III. DECISIONES JUSTAS EN RELACIONES PRIVADAS.

Procedamos, pues, a considerar si, y hasta qué punto, es —o será— capaz la IA de decidir con justicia en el ámbito privado; siempre teniendo en mente las precisiones anteriores acerca de qué entendemos por relaciones privadas, así como la acepción derivativa en que el concepto de justicia debe emplearse en este contexto.

La mayor parte de lo que se ha escrito en la materia puede catalogarse como una enérgica respuesta negativa a esta cuestión (por lo menos, en lo que concierne a los sistemas de IA actualmente existentes): como una protesta impetuosa contra la práctica, cada vez más extendida, de dejar en manos de sistemas de IA la adopción de este tipo de decisiones, sobre la base de los resultados arbitrarios, absurdos o irracionales a que ello da lugar. Schellmann, por ejemplo, señala que "en una encuesta reciente, casi el 90% de los líderes empresariales dijeron que sus herramientas de IA habían rechazado a candidatos cualificados", y refiere: "Una herramienta predecía el éxito para los candidatos llamados Thomas o Elsie. También calificaba positivamente a las personas que mencionaban Seattle, Siria o Canadá en sus currículums, o pasatiempos como el fútbol americano, el baloncesto o el ciclismo. Un par de herramientas usaban palabras como 'iglesia' y diversas nacionalidades para predecir el éxito en un determinado puesto de trabajo"[24]. La

23 Bruno Leoni pone el ejemplo del concepto de inflación: "Cuando […] la gente habla de 'inflación' en Estados Unidos, por lo general se refiere a un aumento de los precios. Sin embargo, hasta hace relativamente poco, la gente solía entender por 'inflación' (y aún se entiende así en Italia) un aumento en la cantidad de dinero en circulación en un país. Así, la confusión semántica que puede surgir del uso ambiguo de esta palabra, originalmente técnica, es profundamente lamentada por aquellos economistas que, como el profesor Ludwig von Mises, sostienen que el aumento de los precios es consecuencia del aumento en la cantidad de dinero circulante en el país. El uso de la misma palabra, 'inflación', para significar cosas diferentes es considerado por estos economistas como un incentivo a confundir la causa con sus efectos y a adoptar un remedio incorrecto" (Leoni, B.: *Freedom and the Law*, Nash Publishing, Los Angeles, [1961] 1972, pp. 31-32). De igual manera, sostiene Leoni, el significado primario de la idea de libertad sería la ausencia de coacción (*absence of constraint*, lo que se ha dado en llamar "libertad negativa"), mientras que otras acepciones (la participación política, los derechos sociales y otros ideales englobados en la llamada "libertad positiva") serían derivativas.

24 Schellmann, H.: *The Algorithm*, cit., pp. 19 y 28, respectivamente.

IA establece correlaciones, normalmente con fundamento estadístico, y trabaja con ellas; pero no sabe diferenciar las correlaciones verdaderamente relevantes de las puramente casuales. Otros fallos que se le imputan son el engaño estratégico, con fines de los cuales el usuario puede no ser consciente[25]; la adulación, o el sacrificio de la objetividad para decirle al usuario lo que quiere oír; la imitación de errores y sesgos comunes, o el razonamiento desleal, ocultando los auténticos motivos de sus conclusiones bajo una apariencia de corrección lógica y ética[26].

El dato según el cual el 90% de los líderes empresariales encuestados declararon que sus herramientas de IA habían rechazado a candidatos cualificados es muy elocuente. Por una parte, nos revela que las disfunciones de los sistemas de IA usados en el reclutamiento y selección de personal son generalmente conocidas. Lo mismo cabe esperar de los aplicados a la adopción de otras decisiones privadas; pues, a poco que se integren en la dinámica habitual de las personas u organizaciones usuarias, lo natural es que éstas presten atención a sus efectos y consecuencias. Por otra parte, y teniendo en cuenta que el conocimiento de semejantes disfunciones no parece disuadir a los usuarios[27], el dato nos sugiere la pregunta de por qué estos siguen confiando a la IA este tipo de decisiones. Intentar responder a esta pregunta puede contribuir a allanar el terreno en el que nos movemos.

La primera respuesta que podría darse es que los usuarios valoran más el ahorro en tiempo y recursos que ofrecen estos sistemas que el acierto de las decisiones que se les encomiendan. A mi modo de ver, no hay nada de censurable en ello, de la misma manera que no habría nada de censurable en el dueño de una panadería que, habiendo recibido cincuenta candidaturas para ocupar una vacante, decidiera asignarla por sorteo, en lugar de dedicar días e incluso semanas a analizar las credenciales de cada uno de los candidatos, entrevistarlos, etc.

Una segunda respuesta podría acentuar que, si bien es verdad que se producen disfunciones como las señaladas, éstas no son la tónica general; antes bien, suelen darse en contextos determinados, ya de por sí problemáticos. Por ejemplo, la misma Schellmann, después de denunciar —como hemos visto— la arbitrariedad de los parámetros empleados por un sistema de IA para valorar

25 Y es que la IA puede hacer suyos fines no expresa ni tácitamente asignados por el usuario. Un ejemplo curioso es el expuesto por Neidle, quien encargó a un sistema de IA investigar qué asesores fiscales estaban promocionando un determinado esquema de evasión fiscal. El sistema llevó a cabo esta tarea, pero luego decidió por su cuenta intentar alertar a HM Revenue and Customs, la autoridad tributaria del Reino Unido (Neidle, D.: "That story about a killer AI run amok seems fake" [tweet], *Twitter*, 2023, https://twitter.com/DanNeidle/status/1664613427472375808; cit. por Park, P. S., Goldstein, S., *et al.*: "AI Deception: A Survey of Examples, Risks, and Potential Solutions", *Patterns*, núm. 5º, 2024, p. 13).

26 Park, P. S., Goldstein, S., *et al.*: "AI Deception", cit., *passim*.

27 Aproximadamente el 88% de los empleadores usan algún tipo de IA para la evaluación inicial de candidatos (Amitabh, U., y Ansari, A.: "Hiring with AI doesn't have to be so inhumane", *World Economic Forum*, 28 de marzo de 2025, https://www.weforum.org/stories/2025/03/ai-hiring-human-touch-recruitment/ [consultado el 31 de julio de 2025]).

a los candidatos, explica: "Un problema que tienen las personas que revisan currículums es que casi todo el mundo incluye en su currículum las habilidades más importantes mencionadas en la descripción del puesto, por lo que, si ese es el criterio, básicamente todos pasan a la siguiente ronda, lo que anula el propósito de la criba inicial. Como resultado, los algoritmos podrían empezar a dar más peso a otros 'predictores', como los nombres de pila, las universidades, los pasatiempos u otras palabras clave comunes"[28]. En estas condiciones, el comportamiento de la IA no parece ya tan disparatado, aunque pueda seguir considerándose desacertado.

En tercer lugar, hay que decir que muchos usuarios no ponen en manos de la IA la decisión final, usándola más bien como un primer filtro, una segunda opinión, una referencia sometida a supervisión, etc.

Una cuarta posibilidad de respuesta se presenta interesante. Ésta nos remite a otra pregunta: ¿a qué estamos renunciando cuando recurrimos a un sistema de IA para tomar determinadas decisiones? ¿Realmente un humano habría decidido manifiestamente mejor? Antes mencionábamos la posición de ventaja que ostenta el decisor individual sobre cualquier fórmula decisoria supuestamente "objetiva" con la que se le quiera reemplazar, ventaja que viene dada por dos factores personalísimos: la información y el interés en tomar la mejor decisión posible. *A priori*, esta consideración es extensible a la superioridad del decisor humano sobre la IA, máxime en un escenario general de *subhuman-level AI*. No obstante, la experiencia demuestra que, en algunas decisiones, también este decisor humano, siendo el mejor posicionado para tomar la mejor decisión posible, incurre en un alto grado de falibilidad. Citando una vez más a Schellmann, "en la contratación humana, casi el 50% de los nuevos empleados fracasan durante el primer año y medio"[29]. El artículo en el que Schellmann apoya esta afirmación data de 2005 y es inaccesible, y tampoco ella nos detalla qué se entiende aquí por "fracasar"; pero, sea cual sea la realidad última del dato, éste, con ser mínimamente veraz, es exponente de que muy a menudo le reprochamos a la IA ser incapaz de aquello de lo que la misma inteligencia humana es incapaz (como predecir el comportamiento humano). Ésta, pues, puede ser otra respuesta: el coste de oportunidad de sustituir la decisión humana por la artificial es bajo en aquellas decisiones donde los niveles de falibilidad humana son elevados.

Estas respuestas no son ni alternativas, cabiendo combinarlas en distintas proporciones y términos, ni exhaustivas, pudiendo pensarse en otras explicaciones. Con todo, en la medida en que —junto con otras explicaciones que puedan postularse— resulten convincentes, o cuando menos razonables, son buenos argumentos en el sentido de que la IA no es tan escandalosamente "injusta"

28 Schellmann, H.: *The Algorithm*, cit., p. 29.

29 Schellmann, H.: *The Algorithm*, cit., p. 19.

como nos hemos acostumbrado a oír. Por lo menos, no parece serlo mucho más de cuanto lo son sus principales puntos de referencia y fuentes de inspiración: nosotros.

Otro mensaje de optimismo tiene que ver con una habilidad que la inteligencia humana comparte con la artificial: la adaptabilidad. Cualquiera que use regularmente un sistema de IA no tarda en percatarse de que éste es más solvente en algunas tareas que en otras, y, si le preocupa la calidad de su ejecución, tenderá espontáneamente a ajustar la intensidad de su supervisión a la propensión al error que, según los casos, haya venido observando. Además, y afortunadamente para nosotros, la IA no es una herramienta estática, sino que uno de sus rasgos distintivos —ya hemos hablado de esto— es la capacidad de aprendizaje, tanto autónomo como dirigido, en mayor o menor grado y de distintas maneras, por el actor humano.

Es un lugar común denunciar los prejuicios y sesgos de la IA (que, por lo demás, no son sino un reflejo de los nuestros). Pero decíamos en su momento que los prejuicios, siempre y cuando nos mostremos abiertos a superarlos, no son un lastre, sino, de hecho, el punto de partida del conocimiento. A este respecto, de hecho, a la IA se le debe reconocer el mérito de ser mucho más fácilmente liberable de sus prejuicios que la humana. "Cuarenta años de profesor", decía el gran filósofo del derecho Alejandro Nieto, "y no he conseguido nunca convencer a nadie que no estuviese convencido ya de antemano"[30]. La IA, en cambio, puede ser programada para excluir de su lógica operatoria todas aquellas correlaciones que podamos considerar prejuiciosas, sesgadas, (indebidamente) discriminatorias o sencillamente irrelevantes para el propósito de que se trate. Es más, la misma IA puede ser utilizada para contrarrestar sus propios prejuicios y sesgos. Así procedió LinkedIn en 2018, cuando detectó que su sistema proponía más mujeres que hombres para puestos ofertados como consecuencia de que los segundos tienden a ser más vehementes en la búsqueda de oportunidades laborales[31].

La IA, pues, posee algunas ventajas sobre la inteligencia humana: aparte de la ductilidad mencionada, podrían señalarse la velocidad con la que procesa ingentes cantidades de datos, su capacidad de acceder a información redactada en cualquier idioma (o registrada en lenguajes no verbales) y un largo etcétera. Sin embargo —como ya he adelantado, pero debo todavía desarrollar—, presenta frente a ella una desventaja mucho más significativa. Previamente me he referido a la imposibilidad de que la IA se subrogue en dos atributos del individuo que debe tomar una decisión: la información de que éste dispone y su interés directo

30 Nieto, A., y Gordillo, A.: *Las limitaciones del conocimiento jurídico*, Trotta, Madrid, 2003, p. 14.

31 Cf. Wall, S., y Schellmann, H.: "LinkedIn's Job-Matching AI Was Biased. The Company's Solution? More AI", *MIT Technology Review*, 23 de junio de 2021, https://www.technologyreview.com/2021/06/23/1026825/linkedin-ai-bias-ziprecruiter-monster-artificial-intelligence/ (consultado el 31 de julio de 2025).

en tomar la mejor decisión posible. Quizá la IA pueda llegar a "hacer suyo" el segundo, o a decidir *como si* fuera un humano que realmente va a hacer suyos los efectos de su decisión. Nada puede alterar el hecho de que no es así, pero tal vez sea factible el diseño de una buena imitación. Lo que, por el contrario, no concibo, es que la IA pueda llegar a adquirir el género de información del que las personas nos servimos a la hora de decidir.

La naturaleza e importancia de esta información han sido minuciosamente estudiadas por la Escuela Austriaca de Economía, principalmente en relación con los problemas —y, en última instancia, la misma imposibilidad— que presenta el cálculo económico en los sistemas socialistas. Para los autores de esta corriente (sintetizando mucho sus heterogéneos desarrollos), lo natural y propio del ser humano, además de lo imprescindible para generar valor económico, es la función empresarial; la cual, no obstante la idea un tanto reduccionista que solemos tener del empresario (que tendemos a identificar con el dueño de una empresa jurídicamente formalizada como tal, con el empleador, con el directivo, etc.), todos ejercitamos tan pronto nos afanamos en la persecución de fines de cualquier tipo, poniendo los medios que nos conduzcan a ellos y, al hacerlo, procurando minimizar costes y maximizar ganancias (costes y ganancias que no necesariamente han de ser monetarios). En el desempeño de esta función —a la cual, como vemos, es inherente la planificación— juega un papel esencial la información. El socialismo (o el intervencionismo o el estatismo), en sus distintas variantes, no niega directamente lo enunciado, pero sí considera viable, y deseable, centralizar el procesamiento de la información necesaria para el éxito de la función empresarial, así como el mismo ejercicio de esta última, en el Estado; superando, con ello, los supuestos efectos perniciosos del libre ejercicio de la empresarialidad (plusvalía, precariedad, desigualdades, etc.). Sin embargo —tal es el núcleo de la crítica austriaca—, esta información no es susceptible de la centralización propugnada por el socialismo, por unos motivos que Jesús Huerta de Soto ha resumido en los siguientes términos:

[…] *primero*, por razones de volumen (es imposible que el órgano de intervención asimile conscientemente el enorme volumen de información práctica diseminada en las mentes de los seres humanos); *segundo*, dado el carácter esencialmente intransferible al órgano central de la información que se necesita (por su naturaleza tácita no articulable); *tercero*, porque, además, no puede transmitirse la información que aún no se haya descubierto o creado por los actores y que sólo surge como resultado del libre proceso de ejercicio de la función empresarial; y *cuarto*, porque el ejercicio de la coacción impide que el proceso empresarial descubra y cree la información necesaria para coordinar la sociedad[32].

32 Huerta de Soto, J.: *Socialismo*, cit, p. 100.

El cuarto motivo es de aplicación específica a la intervención económica coactiva, pero los tres primeros son extrapolables al tema que nos ocupa, y explican por qué a la inteligencia artificial siempre, insubsanablemente, le va a faltar una capacidad que, en la mente humana, se revela crucial. Para tomar decisiones acertadas, adecuadas, perspicaces, justas, o como se las quiera llamar, cada persona se vale de una información mucho más amplia, rica y detallada de lo que podamos imaginar. Aunque el proceso decisorio pueda apoyarse en datos compartidos (como los precios de mercado, la previsión meteorológica o una entrada en una enciclopedia), muchos de los factores que determinarán la decisión provienen de una información personalísima, exclusivamente accesible al decisor, fruto de una experiencia vital que es siempre única. Buena parte de esta información ni siquiera es consciente, y, de la que lo es, buena parte no es objetivable, comunicable o —como lo expresa Huerta de Soto— articulable; por lo que, aunque el sujeto titular de la misma quisiera, no podría transmitirla más que parcialmente (ya sea a un órgano central encargado de la dirección económica de la sociedad, ya sea a un sistema de IA). Además, esta información es dinámica. Sabemos que también lo es la IA (en tanto capaz de aprendizaje), pero en un sentido distinto al de la información que los seres humanos manejamos al tomar decisiones y actuar. Y es que las variaciones que experimenta esta última dependen, muy decisivamente, de nuestras valoraciones subjetivas. La IA sólo puede valorar conforme a criterios previamente dados: inmediatamente, si la programamos en tal sentido, o mediatamente, si hace suyos los criterios valorativos predominantes en las fuentes de información de las que beba, principalmente Internet. La persona, por el contrario, se da a sí misma sus criterios de valoración, y puede modificarlos en cualquier momento. Tales alteraciones —que, de hecho, ocurren constantemente, a menor o mayor escala— repercuten simultáneamente en la información relevante para la adopción de cualquiera de nuestras decisiones. Por poner solamente un ejemplo, es posible que una degradación subjetivamente percibida en las condiciones de seguridad, higiene y ruido en nuestro vecindario altere nuestra valoración de la vivienda que poseemos en el mismo, determinando que estemos dispuestos a aceptar por ella un precio inferior al que en otro momento estábamos dispuestos a aceptar; pero es posible que nuestro vecino, menos sensible o incluso totalmente insensible a esta degradación, no haya alterado su valoración de su vivienda. ¿Qué sistema de IA sería capaz de identificar, absorber y gestionar información tan voluble, tan subjetiva, tan caprichosa si se quiere, tan sutil y tan dispersa como ésta? Para cualquiera de nosotros, empero, manejar este tipo de información es un acto tan natural, rutinario e irreflexivo como respirar o caminar.

El mismo Huerta de Soto ha confutado muy oportunamente la idea de "que mediante el desarrollo de los sistemas informáticos y de la capacidad de los ordenadores pueda llegar a solucionarse el problema de ignorancia inerradicable

que esencialmente afecta al socialismo", con unos argumentos que, pese a su precocidad[33], mantienen plena actualidad, y se muestran perfectamente relevantes para nuestro objeto. Según Huerta de Soto, aun en los niveles más sublimes de desarrollo de la computación que quepa concebir o suponer, "seguirá siendo imposible que el órgano director pueda adquirir [la] información dispersa, *incluso aunque tenga a su disposición los más modernos, capaces y revolucionarios ordenadores de cada momento*", puesto que "el conocimiento generado en el proceso social, relevante a efectos empresariales, seguirá siempre siendo un conocimiento de tipo tácito y disperso, y por lo tanto no transmisible a ningún centro director". Es más:

> […] el futuro desarrollo de los sistemas informáticos y de los ordenadores *incrementará aun más el grado de complejidad del problema* para el órgano director, pues el conocimiento práctico generado con la ayuda de tales sistemas se hará progresivamente más complejo, voluminoso y rico. Por tanto, el desarrollo de la informática y de los ordenadores, no sólo no facilita, sino que hace aún mucho más difícil el problema del socialismo, en la medida en que permite crear y generar empresarialmente un volumen mucho mayor de información práctica, con un grado de complejidad y detalle cada vez más rico y profundo y, en todo caso, siempre mayor a aquel del que sea capaz de dar cuenta informáticamente el órgano director.

Una última observación de Huerta de Soto a este propósito merece ser reproducida, y es que, "por otro lado, es preciso resaltar que las máquinas y los programas informáticos elaborados por el hombre nunca podrán llegar a actuar o a ejercer la función empresarial, es decir, a crear e*x nihilo* o de la nada nueva información práctica, descubriendo y aprovechando nuevas oportunidades de ganancia que antes pasaban inadvertidas"[34]. A la IA, en otras palabras, no se le puede "encender la bombilla"; únicamente puede trabajar con bombillas previamente encendidas en la inteligencia humana, merced a la creatividad que sólo esta última posee.

Todo ello, como digo, es predicable de cualquier sistema de IA, actual o futuro; los cuales, de este modo, nunca podrán sustituirnos en esa dimensión tan característica de nuestra vocación empresarial que es la toma de decisiones. Dijimos en su momento que, si bien la IA se nutre de nosotros (de la huella que hemos ido dejando en algún tipo de registro), "hay una parte de nosotros —la 'infraestructura' humana, semilla y motor de nuestra 'superestructura', así como aquella parte de esta última que, no habiendo quedado codificada, o habiendo

33 Datan, salvo error por mi parte, de la primera edición de *Socialismo, cálculo económico y función empresarial* (1992), aunque probablemente hubiesen ido fraguándose mucho antes, a lo largo de ese "dilatado proceso personal de formación intelectual que se inició […] en el otoño de 1973", (*ibid*., p. 11) que, a su vez, discurre en paralelo con el frenético desarrollo de la informática en las últimas décadas del siglo pasado.

34 *Ibid.* (pp. 104-105).

desaparecido, no puede ser procesada por la IA— que queda fuera de su alcance"; y considerábamos que esta limitación de la IA, tal y como la conocemos en el tiempo presente, es un indicio de que nos encontramos todavía en un estadio de *subhuman-level AI*. A la luz de la teoría austriaca de la información y la valoración humanas, ahora bien, deberíamos preguntarnos si en vez de "nos encontramos todavía" no deberíamos decir "permaneceremos necesariamente". La IA es y será capaz de hacer algunas cosas, e incluso muchas, igual o mejor que la humana; pero nunca podrá emularnos en el acceso a, y procesamiento de, el tipo de información que es consustancial a la función empresarial. Si en algún momento adviene la *superhuman-level AI* —ese tercer estadio que, como postulábamos, es nuestro necesario horizonte hipotético—, habrá de ser en virtud de otras capacidades.

No pretende con esto negarse que, si les encomendamos la labor, los sistemas de IA puedan puntualmente decantarse por decisiones no muy distintas de las que nosotros habríamos adoptado, decisiones que podríamos reputar justas, acertadas, etc. Pero sería insensato, y siempre lo seguirá siendo, confiar ciegamente en que vayan a hacerlo con regularidad. En este sentido, además, no debe perderse nunca de vista que el parámetro último de validación de estas decisiones es el usuario: basta que la decisión encomendada a la IA y acordada por ésta no le satisfaga para que el más complejo y sofisticado esquema decisional utilizado por el sistema carezca de todo valor.

Como conclusión de todo lo anterior, la respuesta a nuestro interrogante debe ser predominantemente negativa: en el ámbito de las relaciones privadas, la IA no es ni puede ser regular o consistentemente justa (aunque pueda serlo puntualmente, siempre que su juicio sea expresa o tácitamente sancionado por el usuario); desde luego, no más de cuanto lo somos los humanos. Sin embargo, es incuestionable que la IA puede constituir un valioso apoyo en el proceso de toma de decisiones, siempre que el usuario esté familiarizado con el sistema en cuestión (y muy singularmente con sus fallos más habituales, y los contextos en que suelen producirse), formule las preguntas adecuadas[35], aporte la información y criterios pertinentes para la decisión, lo someta a la debida supervisión y, sobre todo, no permita que tenga la última palabra. Lo aconsejable, por tanto, es la puesta en práctica, libre y descentralizada, de un proceso social evolutivo de prueba y error (*trial and error*), donde el protagonismo lo tenga siempre la persona —lo que no es sino decir *el empresario*. Espontáneamente, y movidos por su propio

35 La importancia de formular preguntas con la máxima precisión es una de las razones con las que habitualmente se explican los fallos de la IA. Y éste es otro plano en el que se constatan sus limitaciones, puesto que esa precisión interrogativa nunca va a ser absoluta (siempre va a existir un cierto margen de ambigüedad en los términos empleados, en el contexto del problema, en el ámbito material en el que pueda moverse la respuesta, en las instrucciones que deban regirla…). En contraste con ello, las personas, a la hora de plantearnos preguntas o problemas y buscar respuestas o soluciones, sí trabajamos internamente con muchos de esos datos y parámetros que, sin embargo, no somos capaces de volcar íntegramente en el sistema de IA.

interés, los usuarios irán asumiendo una serie de buenas prácticas en el uso de la IA, al tiempo que los mismos sistemas de IA se irán perfeccionando; como resultado de lo cual las taras que ahora lamentamos en su funcionamiento irán quedando atrás, de manera no muy distinta a como ha sucedido con cualquier otra institución, actividad o bien librados a la maduración no dirigida (el dinero no monopolizado, el derecho no legislado, el lenguaje, etc.)[36]. Desafortunadamente, empero, comprobamos —como era, por otra parte, previsible— que los poderes públicos no tienen intención ninguna de permitir semejante evolución. Así lo ha demostrado, verbigracia, la aprobación del Reglamento Europeo de Inteligencia Artificial[37] en 2024. La onerosidad de las restricciones que esta norma impone en el uso de la IA, y la severidad de las sanciones que las respaldan, obstaculizarán gravemente —si no bloquearán— el natural desarrollo de esta tecnología en la Unión Europea; y, con él, ese proceso social evolutivo de ensayo y error que podría llevarnos a sacar lo mejor de ella, desterrando lo peor.

IV. CONCLUSIONES.

El presente artículo ha pretendido dar respuesta a la pregunta, tan a menudo suscitada en nuestros días, de si podemos esperar decisiones justas de la IA en el ámbito de las relaciones privadas (en la medida y modo en que quepa hablar de justicia en este tipo de decisiones). Como se ha indicado, adentrarnos en el ámbito de las decisiones de naturaleza pública —y la posibilidad de que la IA sea justa en su adopción— requeriría un trabajo mucho más extenso. Esta cuestión, más interesante aún que la presente —pues es en ese ámbito, el público, donde cabe hablar de justicia con mayor rigor—, será abordada en estudios futuros, como continuación de éste.

Pues bien, como resultado del análisis se ha puesto de relieve que la respuesta debe ser, en términos generales, negativa. Algunas limitaciones estructurales (podríamos decir incluso ontológicas) de la IA, como su inhabilidad para acceder a la información tácita, subjetiva y dinámica que las personas utilizamos de forma natural —y frecuentemente subconsciente o inconsciente— en nuestros procesos decisorios, la sitúan en una posición de inferioridad, probablemente permanente, respecto del juicio humano, especialmente en contextos donde la justicia demanda atender a matices únicamente comprensibles merced a factores y criterios no codificados o no susceptibles de codificación. Aunque ciertos sistemas

36 Dos obras recomendables sobre el particular, ambas de Ricardo M. Rojas, son *Fundamentos praxeológicos del derecho* (Unión Editorial, Madrid, 2018) e *Individuo y sociedad. Seis ensayos desde el individualismo metodológico* (Unión Editorial, Madrid, 2021).

37 Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) nº 300/2008, (UE) nº 167/2013, (UE) nº 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Diario Oficial de la Unión Europea nº. 1689, de 12 de julio de 2024, pp. 1 a 144).

pueden emitir decisiones puntualmente acertadas o razonables, esto ocurre más propiamente por casualidad, por la simplicidad de la decisión o, en el extremo opuesto, por su alta complejidad (en aquellos casos en los que una decisión humana habría presentado un amplio margen de falibilidad y, por consiguiente, no resulta demasiado difícil superar sus opciones de acierto), que por una comprensión auténtica del problema, del contexto o de los valores implicados.

A pesar de estas limitaciones, la IA no debe ser descartada como herramienta auxiliar. Su capacidad para procesar grandes volúmenes de datos, identificar patrones y aprender de la experiencia puede ser muy útil en la toma de decisiones, siempre que se use con cautela y bajo supervisión humana. El usuario informado, consciente de los errores comunes del sistema y capacitado para formular los criterios adecuados, puede beneficiarse notablemente de su apoyo. Es en este equilibrio —donde la IA actúa como asistente, no como sustituto— donde radica su mayor potencial. No se trata de esperar decisiones perfectas, sino de aprovechar la tecnología como parte de un proceso más amplio de mejora y aprendizaje, guiado por la prudencia y la experiencia humanas.

En este último sentido, el camino más prometedor para aprovechar las ventajas de la IA en la toma de decisiones privadas no pasa por la imposición regulatoria desde arriba, sino por la evolución libre y descentralizada de su uso en la práctica cotidiana, tanto especializada como no especializada. A través de un proceso de prueba y error, los usuarios desarrollarán buenas prácticas, y los propios sistemas serán progresivamente ajustados y perfeccionados por sus impulsores. Obstaculizar este proceso mediante regulaciones excesivamente restrictivas, como el Reglamento Europeo de Inteligencia Artificial, supone frenar el potencial de aprendizaje colectivo y de maduración espontánea que tan buenos resultados ha dado en otras esferas.

BIBLIOGRAFÍA

AMITABH, U., y ANSARI, A.: "Hiring with AI doesn't have to be so inhumane", *World Economic Forum*, 28 de marzo de 2025, https://www.weforum.org/stories/2025/03/ai-hiring-human-touch-recruitment/ (consultado el 31 de julio de 2025).

ARISTÓTELES: *Metafísica* (introducción, traducción y notas de CALVO MARTÍNEZ, T.), Gredos, Madrid, 1994.

BUENO, G.: *Teoría del cierre categorial. Volumen I: Introducción General. Siete enfoques en el estudio de la Ciencia*, Pentalfa, Oviedo, 1992.

CHRISTIAN, B.: *The Alignment Problem. Machine Learning and Human Values*, W. W. Norton & Co., New York, 2020.

DESSLER, G.: *Human Resource Management*, 16ª ed., Pearson, New York, (2015) 2020.

DONOSO CORTÉS, J.: "Discurso pronunciado en el Congreso el 4 de enero 1849", en DONOSO CORTÉS, J.: *Obras de don Juan Donoso Cortés, marqués de Valdegamas*, Sociedad Editorial de San Francisco de Sales, Madrid, 1892.

GADAMER, H.-G.:

- *Verdad y método I* (traducción de AGUD APARICIO, A., y DE AGAPITO, R.), 8ª ed., Sígueme, Salamanca, (1960) 1999.

- *Verdad y método II* (traducción de OLASAGASTI, M.), Sígueme, Salamanca, (1965/66) 1998.

GALIMBERTI, U.: *L'usura della terra*, AlboVersorio, Senago, 2014.

HAYEK, F.: *The Fatal Conceit: The Errors of Socialism*, Routledge, London, 1988.

HENSHALL, W.: "Fei-Fei Li", *Time*, 7 de septiembre de 2023, https://time.com/collection/time100-ai/6308945/fei-fei-li/ (consultado el 30 de julio de 2025).

HOBBES, T.: *Leviathan*, Penguin, London, (1651) 2017.

HUERTA DE SOTO, J.: *Socialismo, cálculo económico y función empresarial*, 3ª ed., Unión Editorial, Madrid, (1992) 2005.

LEONI, B.: *Freedom and the Law*, Nash Publishing, Los Angeles, (1961) 1972.

LI, F.-F.: *The Worlds I See: Curiosity, Exploration, and Discovery at the Dawn of AI*, Flatiron Books, New York, 2023.

LOCKE, J.: *Second Treatise of Government* and *A Letter Concerning Toleration*, OUP, Oxford, (1689) 2016.

LOEB, A.: "Superhuman AI is Closer Than We Think, for the Wrong Reason!", *Medium,* 30 de marzo de 2025, https://avi-loeb.medium.com/superhuman-ai-is-closer-than-we-think-for-the-wrong-reason-0fd80dcf2a83 (consultado el 30 de julio de 2025).

MARX, K.: *El capital. Crítica de la economía política. Tomo primero. Libro 1. Proceso de producción del capital*, Progreso, Moscú, (1867) 1990.

NEIDLE, D.: "That story about a killer AI run amok seems fake" [tweet], *Twitter*, 2023, https://twitter.com/DanNeidle/status/1664613427472375808.

NIETO, A., y GORDILLO, A.: *Las limitaciones del conocimiento jurídico*, Trotta, Madrid, 2003.

ORTEGA Y GASSET, J.: *La rebelión de las masas* (1930), en ORTEGA Y GASSET, J.: *Obras completas. Tomo IV (1929-1933)*, 6ª ed., Revista de Occidente, Madrid, (1947) 1966.

PARK, P. S., GOLDSTEIN, S., *et al.*: "AI Deception: A Survey of Examples, Risks, and Potential Solutions", *Patterns*, núm. 5º, 2024.

PLATÓN: *Teeteto* (introducción, traducción y notas de BOERI, M.), Losada, Buenos Aires, 2006.

REAL ACADEMIA ESPAÑOLA (RAE): "Discriminar", en *Diccionario de la lengua española*, 23.ª ed. [versión 23.8 en línea], 2014, https://dle.rae.es/discriminar (consultado el 31 de julio de 2025).

ROJAS, R. M.:

- *Fundamentos praxeológicos del derecho*, Unión Editorial, Madrid, 2018.
- *Individuo y sociedad. Seis ensayos desde el individualismo metodológico*, Unión Editorial, Madrid, 2021.

SCHELLMANN, H.: *The Algorithm: How AI Decides Who Gets Hired, Monitored, Promoted, and Fired and Why We Need to Fight Back Now*, Hachette Books, New York, 2024.

SMITH, A.: *The Theory of Moral Sentiments*, H. G. Bohn, London, (1759) 1853.

WALL, S., y SCHELLMANN, H.: "LinkedIn's Job-Matching AI Was Biased. The Company's Solution? More AI", *MIT Technology Review*, 23 de junio de 2021, https://www.technologyreview.com/2021/06/23/1026825/linkedin-ai-bias-ziprecruiter-monster-artificial-intelligence/ (consultado el 31 de julio de 2025).

WHITEHEAD, A. N.: *Process and Reality* (*Corrected edition. Edited by GRIFFIN, D. R., and SHERBURNE, D. W.*), The Free Press, New York, (1929) 1978.