REVISTA ELECTRÓNICA
DE ACCESO ABIERTO
ISSN 2282-2313

*Revista Internacional y Comparada de*

# RELACIONES LABORALES Y DERECHO DEL EMPLEO

ADAPT
www.adapt.it
UNIVERSITY PRESS

Revista Internacional y Comparada de RELACIONES LABORALES Y DERECHO DEL EMPLEO

*Escuela Internacional de Alta Formación en Relaciones Laborales y de Trabajo de ADAPT*

# Estudio de la causalidad en la toma de decisiones algorítmicas: el impacto de la IA en el ámbito empresarial*

Adrián ARNAIZ RODRÍGUEZ**
Julio LOSADA CARREÑO***

**RESUMEN:** Una de las principales aplicaciones de la inteligencia artificial (IA) en el ámbito laboral es la denominada "gestión algorítmica", que implica una delegación y ejecución de funciones empresariales en sistemas de IA. Al haberse creado y desarrollado la normativa laboral en torno a personas físicas, y no máquinas, la gestión algorítmica puede originar nuevos problemas a los que no se puede hacer frente con la actual normativa. Uno de estos nuevos problemas incide en el hecho de que estos sistemas basan su funcionamiento en correlaciones (no en causas), exigiendo la normativa laboral la concurrencia de causas (no de correlaciones) en la toma de algunas decisiones empresariales. Esta dicotomía puede generar posibles problemas relativos a la inexistencia y/o falta de suficiencia de las causas legalmente exigidas, o la generación de discriminaciones laborales difíciles de detectar. Por ello, en el presente estudio se abordará el examen de estas cuestiones y la propuesta de diferentes soluciones.

*Palabras clave:* Inteligencia artificial, gestión algorítmica, causalidad, correlación, discriminación algorítmica, existencia y suficiencia causal, derecho del trabajo.



* El presente artículo se ha elaborado en el marco del Convenio entre la entidad pública empresarial Red.es M.P. y la Universidad de Castilla-La Mancha para impulsar la implementación de la Carta de Derechos Digitales en el ámbito de los derechos digitales en el entorno laboral y empresarial C039/23-OT.
** Ingeniero informático; estudiante de Doctorado, ELLIS Alicante (España).
*** Inspector de Trabajo y Seguridad Social.

precontractual. 5.2. Fase contractual. 6. Propuesta de soluciones. 6.1. Soluciones técnicas. 6.2. Soluciones jurídicas. 7. Conclusiones. 8. Bibliografía.

# Studying Causality in Algorithmic Decision Making: the Impact of IA in the Business Environment

**ABSTRACT:** One of the main applications of artificial intelligence (AI) in the workplace is the so-called 'algorithmic management', which involves the delegation and execution of business functions in AI systems. As labour regulations have been created and developed around individuals, not machines, algorithmic management may give rise to new problems that cannot be dealt with under current regulations. One of these new problems is the fact that these systems base their operation on correlations (not on causes), and labour regulations require the concurrence of causes (not correlations) in the making of some business decisions. This dichotomy can generate possible problems related to the inexistence and/or lack of sufficiency of the legally required causes, or the generation of labour discrimination that is difficult to detect. For this reason, this study will examine these issues and propose different solutions.

*Key Words:* Artificial intelligence, algorithmic management, causality, correlation, algorithmic discrimination, causal existence and sufficiency, labour law.

## 1. Introducción

Actualmente estamos viviendo el mayor desarrollo tecnológico de toda la historia de la humanidad, que se ha acelerado exponencialmente en los últimos años, encontrándonos en medio de una revolución industrial que va más allá de la simple automatización de tareas físicas, que fue característica de las tres primeras revoluciones industriales. En estos momentos, las máquinas de automatización cognitiva están adquiriendo un protagonismo creciente, marcando un hito en la evolución tecnológica, pasando las tareas cognitivas de ser ejecutadas por personas a ser desarrolladas por máquinas, transformando la forma en que trabajamos y vivimos. Tal tecnología sería inconcebible hace solo cien años; sin embargo, hoy en día, la hemos normalizado e integrado en diferentes esferas de nuestra vida cotidiana, hasta el punto en que en numerosas ocasiones no somos conscientes de ello, como por ejemplo, es el caso de la detección automática que hace nuestro smartphone de las caras de las personas al ser fotografiadas, la predicción del estado meteorológico de la próxima semana, o el reconocimiento de la matrícula de nuestro vehículo al entrar o salir de un aparcamiento.

Aunque uno de los errores más comunes respecto a la inteligencia artificial (IA) sea no ser conscientes de su presencia y de su uso, también lo es creer que realmente se trata de una tecnología de reciente creación. Lo cierto es que, como ocurre en la mayor parte de los grandes avances de la humanidad, su desarrollo es el resultado del esfuerzo y trabajo de muchas personas durante varias décadas y por ello, parece obvio que el conocimiento de su evolución histórica pueda ser un buen comienzo para su estudio, y un reconocimiento a la dedicación que a la misma han hecho tantos profesionales.

## 2. Breve historia de la IA y de su evolución

Desde tiempos inmemorables la humanidad ha soñado con dotar a las máquinas de la capacidad de razonar y tomar decisiones de manera análoga a los seres humanos. Este deseo empezó a ser una realidad a mediados del siglo pasado, ramificándose el camino para conseguirlo en diversas corrientes de pensamiento que han dado forma a la evolución de los sistemas de IA a lo largo de las últimas décadas, llegando a alcanzar un grado de madurez que permite usarla con fines comerciales.

Por ello, su desarrollo ha pasado por distintas etapas, tanto desde el punto de vista teórico como práctico, no teniendo ni mucho menos una

evolución lineal y progresiva, experimentando algunas ralentizaciones denominadas inviernos de la IA (*AI Winters*). Concretamente, existieron 3 etapas separadas por la existencia de dos inviernos que supusieron una desaceleración de los avances científicos con la consiguiente congelación de las inversiones, a cuyo sucinto estudio nos referiremos a continuación[1].

## 2.1. IA fundacional (1950s-1970s)

En primer lugar, es necesario destacar que el nacimiento de la IA no puede entenderse prescindiendo de las contribuciones que a esta materia realizó el matemático Alan Mathison Turing, que ha sido calificado por muchos como el padre de ésta. Una de sus principales contribuciones fue el *Manifiesto Turing*, que sirvió para asentar los principios teóricos que posteriormente inspirarían a generaciones de investigadores, constituyendo el nacimiento de dos escuelas de pensamiento: la escuela simbolista de enfoque *top-down*, y la escuela conexionista de enfoque *bottom-up*. Cada uno de estos paradigmas ha influido de manera significativa en cómo se concibe y diseña la IA en la actualidad.

Posteriormente, Frank J. Rosenblat inventó el "perceptrón", una neurona artificial, inspirada en las neuronas biológicas del cerebro humano, que se constituiría como la unidad básica de las redes neuronales. Este hito, entre muchos otros, dio lugar a un optimismo desmesurado que fue seguido de atrevidas predicciones, que finalmente no se materializaron, dando lugar a lo que se conoce como el primer invierno de la IA. Entre los principales obstáculos que generaron este estancamiento se pueden destacar las limitaciones en la capacidad de computación, memoria y velocidad de procesamiento.

## 2.2. IA simbólica (1970s-1990s)

En la década de 1980, de mano de investigadores como Allen Newell y Herbert A. Simon, empezó a adquirir relevancia el denominado enfoque simbólico, que se basaba en la creencia de que es posible describir la complejidad del mundo, y los factores que en él intervienen, mediante el

[1] B. DELIPETREV, C. TSINARAKI, U. KOSTIĆ, *AI Watch. Historical Evolution of Artificial Intelligence. Analysis of the three main paradigm shifts in AI*, JRC Technical Report, 2020, pp. 3-8.

uso de un lenguaje formal que los ordenadores pudiesen entender[2]. Este enfoque se basaba en la creencia de que las capacidades cognitivas superiores podrían ser alcanzadas mediante la programación de reglas simbólicas complejas, basando su funcionamiento en la sentencia informática de "si pasa esto, haz lo otro". Por ello, en el enfoque simbólico se efectúa una previa programación de todas y cada una de las situaciones a las que el sistema se va a enfrentar y todas y cada una las respuestas que el sistema debería dar, lo que genera que estos sistemas tengan un carácter determinista, lo cual derivaría en el hecho de que para una misma entrada de datos siempre se daría una misma salida.

Sin embargo, al ser sistemas que estaban basados en el previo conocimiento humano, su desarrollo estaba limitado por la adquisición y actualización de este, circunstancia que se conjugó con su imposibilidad técnica para resolver problemas genéricos (ya que solo podían solucionar los concretos y específicos para los cuales habían sido previamente programados), lo que finalmente derivó en que las expectativas existentes no fueran materializadas[3]. Esta nueva desaceleración en las innovaciones científicas en la materia se conocería como el segundo invierno de la IA, y daría lugar a la necesaria exploración de alternativas científicas.

## 2.3. IA subsimbólica (1990s-2020s)

Durante la década de 1990, como fruto de la aparición de internet, la generación e intercambio de datos creció drásticamente, lo que en combinación con un crecimiento exponencial de la capacidad de cómputo de los procesadores y el desarrollo de nuevos algoritmos, como el de *backpropagation*[4], creó el ecosistema perfecto para propiciar un cambio de enfoque en el desarrollo e investigación de la IA, pasando de un enfoque simbólico a un enfoque conexionista o subsimbólico.

En este último enfoque, y a diferencia del anterior, los desarrolladores del sistema de IA no efectúan una previa programación para hacer frente a todas y cada una de las situaciones que pueden acontecer, sino que, habiendo codificado unas reglas básicas, este sistema aprende a resolver una o varias tareas mediante los datos que le son proporcionados, pudiendo posteriormente adaptarse al entorno mediante la experiencia derivada de la

[2] A. CHRISTENKO *ET AL.*, *Artificial intelligence for worker management: an overview. Report*, EU-OSHA, 2022, p. 11.
[3] R. LÓPEZ DE MÁNTARAS BADIA, P. MESEGUER GONZÁLEZ, *Inteligencia artificial*, Catarata, 2017, pp. 50-51.
[4] I. GOODFELLOW, Y. BENGIO, A. COURVILLE, *Deep Learning*, MIT Press, 2016.

interacción con el mundo real.

La escuela subsimbólica se basa en un enfoque *bottom-up* (de abajo hacia arriba), es decir, que a partir de los datos se obtenía un conocimiento presente (y oculto) en los mismos (los denominados *Insights*). Por lo tanto, los datos adquieren una gran relevancia en su funcionamiento, pues son estos los que permiten que el sistema se adapte a cada situación y halle la mejor forma para alcanzar el objetivo para el que previamente ha sido programado, sin que haya un procedimiento concreto previamente establecido por los desarrolladores de este (lo que supone una cierta autonomía en la *praxis* de la máquina, pero no en la fijación de sus objetivos)[5]. La consecuencia de todo ello es un comportamiento no determinista, probabilístico o estocástico, que implica una incertidumbre sobre los datos de salida que va a producir el sistema, y, por ende, de su interacción con el ambiente.

Finalmente, es importante destacar que uno de los principales hitos de la escuela conexionista es el desarrollo de la categoría científica del aprendizaje de máquina (*Machine Learning* – ML), y especialmente de su subcategoría de aprendizaje profundo (*Deep Learning* – DL). Esta última se caracteriza por utilizar unos algoritmos llamados redes neuronales profundas (*Deep Neural Networks* – DNN), que incorporan múltiples capas de procesamiento, y que demuestran una capacidad excepcional para aprender representaciones jerárquicas de datos. A pesar de sus ventajas, los sistemas de IA basados en DNN presentan desafíos significativos derivados de algunas de sus peculiaridades técnicas. De entre ellas, y con especial incidencia en el ámbito laboral, cabe destacar las siguientes[6]:

- un funcionamiento basado en la detección de correlaciones y patrones en los conjuntos de datos con los que son entrenados, lo que *a priori* supone una clara dificultad en la detección de una causa legal en la toma de sus decisiones, con la correspondiente posible nulidad de la decisión algorítmica por inexistencia o insuficiencia de causa (cuando esta es exigida legalmente), o la posible generación de discriminaciones laborales de difícil detección;
- el llamado efecto "*black box*" o "caja negra", que alude a la falta de transparencia y claridad en cómo estos sistemas toman sus decisiones, lo que en materia laboral dificulta que la persona

[5] S. TOLAN *ET AL*., *Measuring the Occupational Impact of AI: Tasks, Cognitive Abilities and AI Benchmark*, European Commission, 2020, pp. 191-193.

[6] *Cfr.* M. KEARNS, A. ROTH, *The Ethical Algorithm. The Science of Socially Aware Algorithm Design*, Oxford University Press, 2019; N. OLIVER, *Governance in the era of data-driven decision-making algorithms*, en A. GONZÁLEZ, M. JANSEN (eds.), *Women Shaping Global Economic Governance*, CEPR Press, 2019.

trabajadora destinataria de una decisión algorítmica pueda obtener una explicación real y entendible de por qué se ha adoptado la misma. Por ello, en el ámbito laboral se ha puesto énfasis en la necesidad de transparencia, que se intenta hacer efectiva mediante la llamada información algorítmica.

Ambas características pueden colisionar con el Derecho del Trabajo, generando posibles perjuicios a las personas trabajadoras, que son difíciles de resolver con la actual normativa. Es esta circunstancia la que motiva a desarrollar el presente estudio, centrándonos en el uso empresarial de los sistemas de IA.

## 3. Uso de la IA en el ámbito empresarial

En primer lugar, conviene recordar que durante los últimos tres siglos la productividad y eficiencia de la fuerza laboral ha experimentado un crecimiento exponencial, fruto de la creación e implementación de diferentes tecnologías en los procesos productivos de bienes y servicios. El desarrollo de estas tecnologías se agrupa comúnmente en cuatro revoluciones industriales, basándose las tres primeras en la automatización o semiautomatización de tareas físicas, y la cuarta en la automatización o semiautomatización de tareas cognitivas mediante la IA.

Debido a que el Derecho del Trabajo se crea y desarrolla en torno a estas revoluciones industriales, teniendo por objetivo la protección de los derechos e intereses de la parte contractual más débil, la parte trabajadora, parece indispensable, que el *ius laboralista* aprecie las peculiares características que estas tecnologías manifiestan en los procesos productivos empresariales, ya que pueden generar riesgos y problemas de diferente índole para las personas trabajadoras, lo que en última instancia debería desembocar en la aprobación de nuevas normas o en la actualización de las ya vigentes.

Este objetivo debe desarrollarse eficazmente, pero sin afectar negativamente a la competitividad e innovación empresarial, ya que en numerosas ocasiones los intereses de ambas partes no son convergentes, generando un conflicto de intereses que adquiere una nueva dimensión con la incorporación de la IA en los procesos productivos empresariales, y que a los efectos que aquí nos atañen, pueden ser condesados en una serie de ventajas y riesgos derivados de su uso.

### 3.1. Ventajas y riesgos del uso de la IA en la toma de decisiones

En un mundo cada vez más globalizado y competitivo, la búsqueda de alguna ventaja diferencial se erige como uno de los objetivos más deseados en todo el sector empresarial. En este sentido, la madurez tecnológica alcanzada por la IA en general, y por el ML en particular, hace posible su uso comercial por las diferentes empresas, otorgando en muchos casos una ventaja competitiva determinante, al tener la capacidad de analizar y detectar patrones en grandes cantidades de datos, y permitir la adopción de la decisión óptima en cada momento. De hecho, el acceso masivo a flujos de datos sobre el comportamiento humano y la aplicación de técnicas analíticas mejoradas, principalmente a través de técnicas de ML, está permitiendo a las empresas resolver problemas de gran complejidad, cuya resolución no sería posible de otra forma[7].

Por añadidura, históricamente, los humanos no hemos tomado decisiones perfectas debido a una combinación de diversas razones, como son, la incapacidad de analizar el problema completo, los conflictos de intereses, la corrupción, el egoísmo o los sesgos cognitivos, lo cual ha desembocado en muchas ocasiones en resultados y decisiones injustas o subóptimas[8]. Ante esta situación, diferentes expertos cualificados en IA han propugnado el uso de algoritmos como solución para superar estos problemas, permitiendo alcanzar mejores decisiones con criterios más objetivos y actualizados en tiempo real[9]. Es necesario indicar que, aunque generalmente los sistemas de IA no están aquejados de los sesgos propios del comportamiento humano, a veces pueden aprender comportamientos discriminatorios que están presentes en los datos con los que son entrenados, generando o amplificando las mismas y contraviniendo la legalidad vigente[10].

Como los sistemas de IA pueden emplearse para automatizar o semiautomatizar casi cualquier tarea humana, estas ventajas e inconvenientes pueden generarse en todas aquellas tareas desarrolladas por un sistema de IA en el seno empresarial, ya se encuadren dentro de las facultades empresariales de gestión de las personas trabajadoras, ya se enmarquen en las tareas propias realizadas por estas últimas. Más

[7] M. WILLSON, *Algorithms (and the) everyday*, en *Information, Communication & Society*, 2017, vol. 20, n. 1.
[8] S.T. FISKE, *Stereotyping, Prejudice, and Discrimination*, en D.T. GILBERT, S.T. FISKE, G. LINDZEY (eds.), *The Handbook of Social Psychology. Volume One*, McGraw-Hill, 1998.
[9] N. OLIVER, *op. cit.*
[10] M.L. RODRÍGUEZ FERNÁNDEZ, *Inteligencia artificial, género y trabajo*, en *Temas Laborales*, 2024, n. 171.

concretamente, cabe diferenciar entre:

- las tareas que desarrolla el empresario, o personas que actúen por delegación de éste (los cuadros intermedios, directivos o "encargados"), y que comúnmente materializan las funciones de dirección/organización, vigilancia/control y recompensa/sanción que están reconocidas en la normativa laboral, siendo todas ellas, tareas de naturaleza cognitiva. Por ejemplo, las decisiones de contratar y despedir a una u otra persona, o asignar determinados turnos o tareas, etc.;
- las tareas que desarrollan los trabajadores como parte del ciclo productivo de la empresa, que previamente han sido asignadas por el empresario o por el encargado. Estas tareas pueden tener naturaleza física (como realizar funciones de peón en una obra) o cognitiva (como conceder créditos, desarrollar funciones de abogacía en litigios, o efectuar diagnósticos y tratamientos médicos).

Todas las tareas cognitivas anteriormente expuestas tienen naturaleza decisoria y aunque tradicionalmente han sido desarrolladas por seres humanos, progresivamente están siendo atribuidas a sistemas de IA por motivos de eficiencia económica. No obstante, los riesgos asociados con la implementación de estos sistemas varían dependiendo de si se automatiza o semiautomatiza una tarea de gestión de personal o una tarea relacionada con el ciclo productivo de la empresa. Dicho con otras palabras, cuando se utilizan sistemas de IA para la gestión de personal, el elemento nuclear que genera la asimetría de poder en las relaciones laborales a favor de la empresa ya no se desarrolla por un ser humano, sino por una máquina. Esto genera riesgos significativos para los intereses de las personas trabajadoras que la normativa actual no puede abordar adecuadamente.

Por consiguiente, en este contexto, se impone ser muy precavidos a la hora de implementar sistemas de IA para la toma de decisiones laborales, siendo esta la razón por la que nos centraremos únicamente en los sistemas de IA que son utilizados para gestionar a las personas trabajadoras (la llamada "gestión algorítmica"), y dejaremos al margen de nuestro estudio los sistemas de IA que se utilizan para automatizar o semiautomatizar tareas del ciclo productivo, aunque tengan naturaleza decisoria.

En este trabajo nos centraremos en el estudio de la existencia y suficiencia de causa en las decisiones empresariales que hayan sido adoptadas por los sistemas de IA previa delegación empresarial en los mismos, enfocando esta cuestión en una doble dimensión: la dimensión positiva (esto es, concurrencia de causa legal) y la dimensión negativa (esto es, ausencia de causa discriminatoria). Para abordar el estudio de la causalidad en la toma de decisiones algorítmicas procederemos a explicar

brevemente el concepto de sistemas de IA y su regulación jurídica, para posteriormente exponer los conceptos de causalidad, correlación y efecto *black box*, lo que nos permitirá comprender su funcionamiento y, por ende, abordar con una cierta profundidad las dos dimensiones del problema aquí examinado.

## 4. Concepto de sistema de IA. Aspectos jurídicos. Aspectos técnicos

### 4.1. Concepto de sistema de IA

Para comenzar, es necesario destacar que el concepto de IA no hace referencia a una tecnología única y singular, sino a una disciplina científica cuyo objetivo común es lograr una abstracción matemática de los procesos intelectuales y cognitivos del cerebro humano. La gran complejidad de este ambicioso objetivo dificulta enormemente la función legislativa para establecer una definición unitaria de la misma. A pesar de ello, el Reglamento IA[11] establece una definición cuyo tenor literal es el siguiente:

> Sistema de IA: un sistema basado en máquinas diseñado para funcionar con diversos niveles de autonomía y que puede mostrar capacidad de adaptación tras su despliegue y que, para objetivos explícitos o implícitos, infiere, a partir de la entrada que recibe, cómo generar salidas tales como predicciones, contenidos, recomendaciones o decisiones que pueden influir en entornos físicos o virtuales.

Además, debe indicarse que se trata de una disciplina científica en constante evolución, habiéndose acelerado la misma exponencialmente en los últimos años, razón por la que es complicado exponer de forma estructurada y duradera las ideas que el jurista pueda tener sobre la misma. No obstante, puede ser útil en el estudio de esta materia saber que los sistemas de IA tradicionalmente se clasifican, atendiendo a la amplitud de las funciones que pueden desarrollar, en IA débil (*Artificial Narrox Intelligence* – ANI), IA fuerte (*Artificial General Intelligence* – AGI), e IA superinteligente (*Artificial SuperIntelligence* – ASI). Cada uno de estos tipos de IA sería

[11] 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.

coincidente con una época histórica en el desarrollo de esta tecnología, encontrándonos actualmente en una etapa intermedia entre el desarrollo de la ANI y de la AGI. Uno de los últimos productos comerciales lanzados al mercado, fruto de esta transición, son los llamados "Sistemas de IA de propósito general" (*General Purpose AI* – GPAI) como GPT-4 o Gemini. Los GPAI se encuadran dentro de la llamada IA generativa, caracterizándose por su multimodalidad, teniendo la capacidad de recibir y proporcionar diferentes tipos de datos (texto, audio, imagen) y adaptarse a diferentes tipos de tareas, actuando como un asistente al controlador humano.

A raíz de esta innovación tecnológica, los sistemas de IA han comenzado a clasificarse recientemente en dos categorías que facilitan el desarrollo de funciones específicas y ofrecen diferentes ventajas comerciales, la IA predictiva y la IA generativa:

- los sistemas de *IA predictiva* se utilizan para analizar patrones históricos y actuales con el objetivo de hacer predicciones fundamentadas. Su funcionamiento se basa en algoritmos estadísticos y en el ML. Su principal beneficio comercial radica en proporcionar información valiosa para la toma de decisiones en diversas áreas, desde los negocios hasta la atención médica. La IA predictiva puede ofrecer diferentes tipos de información de salida, pudiendo ser esta a) una predicción o una recomendación, lo cual conllevará una semiautomatización de la tarea, pues el sistema ofrece una asistencia al controlador humano, y b) una decisión[12], lo cual implicará una automatización de la tarea, pues el sistema ejecuta por sí mismo una acción, sin perjuicio de una posible revisión posterior por parte del controlador humano;
- los sistemas de *IA generativa* se utilizan para crear contenido nuevo y original. Estos utilizan el DL para generar contenido basado en los datos con los que han sido entrenados. Pueden ser utilizados en áreas como el arte, el diseño, la música y la escritura creativa. La IA generativa proporciona únicamente contenidos como datos de salida.

*A priori*, y sin perjuicio de futuros desarrollos técnicos, solo la IA predictiva se utiliza para desarrollar funciones de gestión de las personas trabajadoras, pero limitamos nuestro estudio a cómo la misma interacciona

[12] Por ejemplo, un sistema de predicción podría estimar el rendimiento de un determinado trabajador, un sistema de recomendación podría predecir dicho rendimiento y proponer al controlador humano la toma de una decisión concreta como abonarle un plus de productividad o sancionarle, y un sistema de decisión automatizada podría ejecutar las acciones citadas sin una previa intervención humana.

con el Derecho del Trabajo.

### 4.2. Aspectos jurídicos de los sistemas de IA en la gestión algorítmica

Tal como anteriormente se ha indicado, esta tecnología tiene una implantación y un uso transversal a los diferentes ámbitos sociales y productivos, generando en todos ellos nuevos riesgos e incrementando algunos ya existentes. Por ello, la normativa europea reguladora de IA es elaborada con un enfoque basado en riesgos, diferenciando una serie de prácticas prohibidas y tres niveles de riesgo (alto riesgo, riesgo limitado y riesgo mínimo), que conllevan la asignación de diferentes obligaciones jurídicas. Aunque la normativa europea califica de alto riesgo algunos de los sistemas de IA que pueden ser utilizados para la gestión algorítmica, las obligaciones derivadas del Reglamento IA no interfieren en el objeto del presente estudio, y, por lo tanto, no serán aquí estudiadas.

Además, y en lo que respecta a la gestión algorítmica suele tenerse en cuenta el posible tratamiento de datos personales con la consiguiente aplicación del Reglamento general de protección de datos (RGPD)[13], y más concretamente, lo dispuesto en el art. 22 de la citada norma, que es relativo a la toma de decisiones automatizadas. No obstante, el presente trabajo no tiene por objeto estudiar posibles riesgos en materia de privacidad de las personas trabajadoras o examinar en qué supuestos se pueden automatizar decisiones empresariales, sino exponer el problema de la posible dificultad (o incluso imposibilidad) de detectar cuál es la causa subyacente a la automatización o semiautomatización de una decisión de gestión de personal. Por tanto, tampoco entraremos en el estudio de estas cuestiones jurídicas.

### 4.3. Aspectos técnicos de los sistemas de IA en la gestión algorítmica

En el contexto de la gestión de personas trabajadoras mediante sistemas de IA, es imperativo comprender los desafíos fundamentales que surgen de su funcionamiento basado en correlaciones y no en causas. Para

[13] Reglamento (UE) 2016/679 del Parlamento europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos.

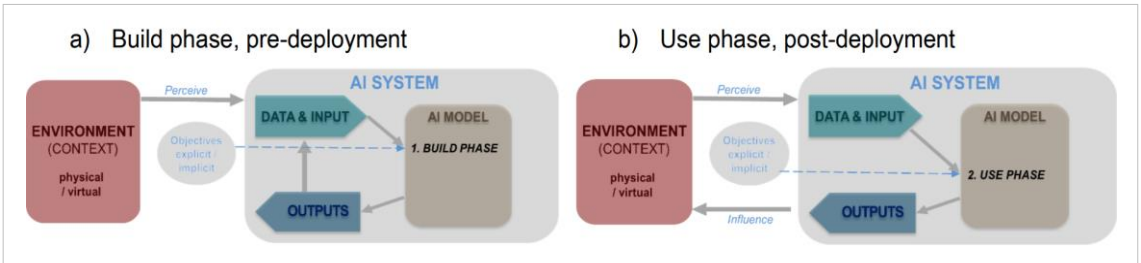
ello, en este análisis se explorarán los aspectos técnicos y conceptuales sobre cómo los modelos de IA aprenden de correlaciones, y de por qué la causalidad suele ser difícil de determinar. Por consiguiente, vamos a exponer brevemente el funcionamiento de un sistema de IA, para posteriormente desglosar los conceptos de correlación, causalidad y *black box*, explorando sus posibles relaciones con la generación de discriminaciones algorítmicas.

### 4.3.1. Funcionamiento de sistemas de IA en la toma de decisiones

En primer lugar, conviene recordar que los sistemas de IA son programas de software que actúan con un cierto grado de autonomía para alcanzar los objetivos para los que han sido programados, obteniendo datos (estructurados o no) del ambiente con el que interactúan, a través de los sensores de los que disponen. Ulteriormente, procesan la información obtenida, ya sea usando las reglas previamente codificadas (enfoque simbólico) o mediante un modelo matemático de detección de patrones que han aprendido, resultante del entrenamiento al que han sido sometidas (enfoque subsimbólico). Por último, interactúan con el ambiente a través de los denominados "actuadores", ya sean estos físicos (como unos brazos robóticos) o digitales (como la pantalla en la que se muestra un texto generado por un GPAI)[14]. Los sistemas de IA, como cualquier máquina, tienen un ciclo de vida dividido en dos grandes fases: pre-implantación (desarrollo y entrenamiento del sistema de IA) y post-implantación (uso del sistema de IA). En ambas fases existe una fuerte dependencia hacia los datos, pudiendo gráficamente ser representadas de la siguiente forma[15].

[14] HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, *A Definition of AI: Main Capabilities and Scientific Disciplines*, 2019.
[15] OECD, *Explanatory memorandum on the updated OECD definition of an AI system*, OECD Artificial Intelligence Paper, 2024, n. 8.

**Figura 1** – Sistema de IA




**Fuente**: OECD, *Explanatory memorandum on the updated OECD definition of an AI system*, OECD Artificial Intelligence Paper, 2024, n. 8, p. 7, figura 1

En la actualidad, los sistemas de IA más comunes son los denominados sistemas de IA de enfoque subsimbólico (como los basados en ML), cuyo funcionamiento se basa en el entrenamiento del sistema mediante ingentes cantidades de datos, los denominados "datos de entrenamiento" (*Dataset*). El objetivo es encontrar las correlaciones y aprender los patrones ocultos que hay en ellos, lo que les permitirá predecir algunas características de futuros datos, y realizar una o varias tareas concretas. Cuando el sistema de IA ha hallado las correlaciones y patrones entre los diferentes conjuntos de datos se le denomina generalmente "modelo de IA" (*AI model*)[16].

Sin embargo, siguiendo la definición de la OCDE, debemos considerar el sistema de IA como un todo, un (sistema) conjunto completo, que puede englobar el modelo de IA y capas adicionales de procesamiento. De hecho, es importante destacar que un sistema de IA (*AI system*), al ser un software complejo, puede constar de varias capas o componentes que se adicionan al modelo de IA, aunque es este último el que actúa como el verdadero corazón del software[17]. Por consiguiente, los datos de salida que proporciona el modelo de IA pueden ser recibidos directamente por el controlador humano o ser procesados por otros componentes del sistema para posteriormente ser suministrados al usuario. Dicho con otras palabras, aunque el corazón del sistema sea un modelo de IA, los datos de salida de

[16] S. BAROCAS, A.D. SELBST, *Big Data's Disparate Impact*, en *California Law Review*, 2016, vol. 104, n. 3, p. 677.

[17] El sistema de IA es como una alcachofa, que tiene un "corazón" y varias capas. Generalmente, la parte más útil de la alcachofa es su corazón (en el sistema de IA es el modelo de IA), el cual está rodeado por varias capas que envuelven al mismo. Por lo tanto, aunque en muchas ocasiones se utilizan como sinónimos, no son exactamente coincidentes los significados de los conceptos de sistema de IA (que se refiere a todo el conjunto) y de modelo de IA (que se refiere al corazón de este), siendo este último el que en puridad aprende las correlaciones existentes en los conjuntos de datos.

este modelo pueden ser los de entrada para otro componente que aplique reglas o algoritmos clásicos (como los de enfoque simbólico) para presentar o ejecutar la información de una manera más adecuada y exacta. Este enfoque permite que el sistema aproveche tanto la capacidad predictiva de la IA como la capacidad de procesamiento específica de otros tipos de software, que permite no solo predecir, sino también recomendar o automatizar decisiones. En efecto, en puridad un modelo de IA solo puede ofrecer como datos de salida predicciones probabilísticas, que serán consideradas recomendaciones o predicciones dependiendo de cómo se gestionen en las capas adicionales de procesamiento. El conocimiento de estas vicisitudes adquirirá especial relevancia en la diferenciación entre la semiautomatización (cuando los datos de salida son predicciones o recomendaciones) y la automatización de las tareas laborales (cuando los datos de salida son decisiones).

En conclusión, y a los efectos que aquí nos atañen, debe indicarse que es el modelo de IA el que infiere las relaciones entre variables basándose en estadística, careciendo, sin embargo, de capacidad para establecer relaciones directas de causa y efecto. Esta limitación técnica puede generar predicciones, recomendaciones o decisiones no basadas en relaciones causales (que, además, en ocasiones pueden estar sesgadas), lo que, en última instancia implica la posible nulidad de las decisiones empresariales que sean delegadas en los mismos, debido principalmente a una ausencia o insuficiencia de causa legal. Debido a la complejidad técnica de estas afirmaciones, se hace necesario explicar los conceptos de correlación y causalidad.

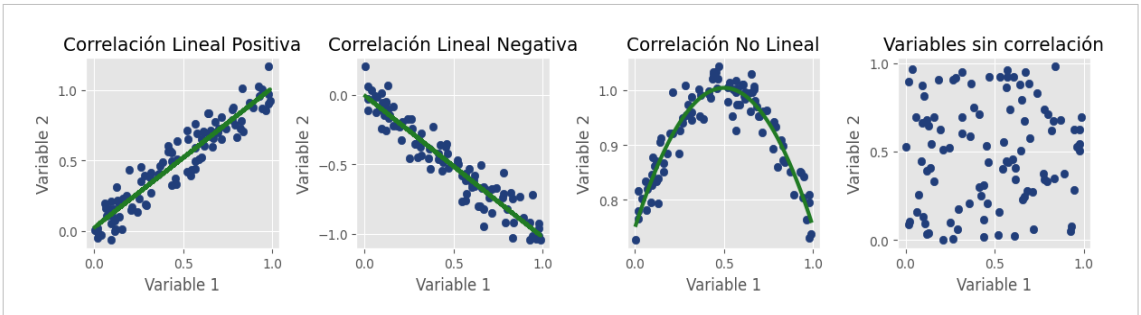
### 4.3.2. Explicación del dilema correlación-causalidad. Ejemplos en el ámbito laboral

En primer lugar, la correlación, en su forma más simple, es una relación estadística entre dos variables que tienden a cambiar juntas, lo que supone que la correlación refleja una relación de dependencia entre dos elementos. Si conociendo el valor de una variable es posible saber o estimar el valor de una segunda, ambas variables están relacionadas o correlacionadas. La correlación nos permite saber si existe o no un patrón estadístico entre dos variables, pero no permite constatar una relación de causalidad entre ambas[18]. Esta relación de correlación entre dos variables puede ser lineal (cuando una sube o baja, la otra sube o baja), o no lineal (cuando la relación

[18] J. PEARL, *Causality. Models, Reasoning, and Inference*, Cambridge University Press, 2009.

sigue un patrón más complejo).

**Gráfico 1** – Tipos de correlación entre variables




Esta dependencia se puede medir técnicamente con varias herramientas, métodos y conceptos de estadística y probabilidad. Además, la mayoría de los métodos actuales de toma decisiones están basados en explotar estas correlaciones para encontrar patrones subyacentes en los datos, de tal forma que ayuden a la creación de predicciones, sin que, por ello, el sistema de IA tenga capacidad para hallar relaciones causales[19].

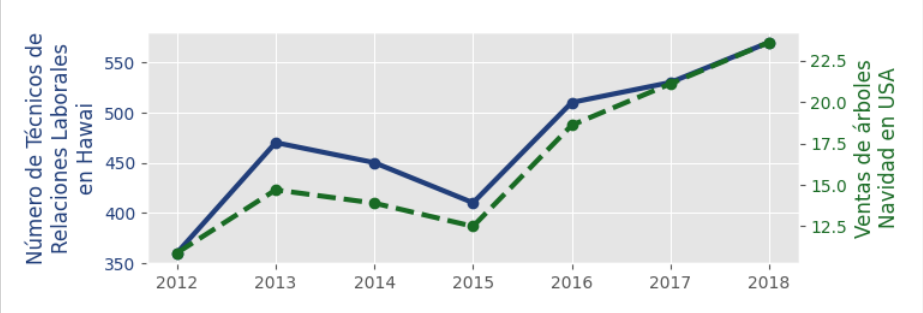
De hecho, que dos variables o conjuntos de variables estén correlacionadas no implica necesariamente una relación causa-efecto entre ellas. Ni la correlación indica si hay causalidad, ni tampoco en qué dirección se produce, es decir, no se sabe cuál es la causa y cuál el efecto, solo se sabe que ambas variables cambian a la vez. En otras palabras, aunque dos variables puedan estar correlacionadas, esto no significa que un cambio en una variable cause un cambio en la otra. Esta distinción es crucial en el ámbito laboral, donde la existencia y suficiencia de causa se erige en numerosas ocasiones como un elemento esencial en la toma de decisiones empresariales, ya sean tomadas por seres humanos, ya sean delegadas en sistemas de IA, y cuya inexistencia o insuficiencia tiene importantes efectos jurídicos, como la declaración de nulidad de estas y la posible imposición de una sanción administrativa.

Por tanto, es crucial ahondar en el estudio del concepto de la correlación y comprender los factores que pueden originarla, que son los tres siguientes.

[19] J. PETERS, D. JANZING, B. SCHÖLKOPF, *Elements of Causal Inference. Foundations and Learning Algorithms*, MIT Press, 2017.

*1) Por pura aleatoriedad y casualidad (generándose la denominada "correlación espuria")*

**Gráfico 2** – Ejemplo de correlación espuria entre el número de técnicos de relaciones laborales en Hawai y las ventas de árboles de navidad en USA (en millones)




**Fuente**: elaboración propia sobre datos de T. VIGEN, *Spurious correlations. Correlation is not causation*, en *tylervigen.com*, 14 mayo 2015

Las correlaciones espurias se basan en una mera aleatoriedad, no habiendo ninguna relación causal directa entre las variables, ni ningún factor externo que las relacione. Un ejemplo gráfico puede ser el mostrado en el Gráfico de arriba.

Este tipo de correlación refleja lo que un humano percibe como casualidades (no causalidades), y carece de un verdadero valor en la toma de decisiones, pues detecta simplemente hechos que podrían ser calificados como "curiosos".

*2) Por causalidad directa*

Este segundo tipo de correlación implica una relación directa de causa y efecto entre variables, de tal forma que en una de ellas provoca necesariamente un cambio en la otra, pudiéndose conocer de antemano las consecuencias de dicho cambio. Habitualmente se expresa como "si pasa X, entonces ocurre Z". Para hallar estas relaciones de causalidad es necesario tener una comprensión profunda de los mecanismos subyacentes y la capacidad de demostrar que no existen factores externos no identificados previamente.

La correlación puede sugerir una posible relación causal, pero no la garantiza. Por ello es de notable importancia que estas relaciones estadísticas sean interpretadas por una persona con conocimiento del dominio afectado. Porque identificar relaciones causales requiere un análisis

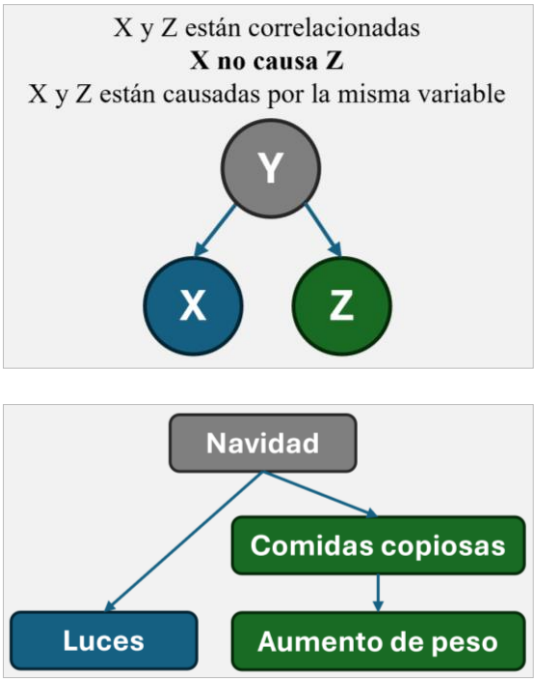
más profundo y la consideración de variables externas que puedan explicar la correlación sin una relación de causa y efecto entre ellas[20].

*3) Por factores externos no identificados*

En este caso la correlación entre dos variables puede surgir debido a la influencia de una o más variables externas, conocidas como variables de confusión. Estas variables externas pueden afectar a las ya examinadas de varias maneras, creando así una aparente correlación entre ellas. En relación con esta cuestión es necesario indicar que existen, a efectos laborales, principalmente dos supuestos en los que dos variables pueden estar correlacionadas sin existir una relación de causalidad[21].

*1) Ambas variables tienen una causa común*

**Figuras 2 y 3** – Dos variables correlacionadas debido a una variable común




En este supuesto, una misma variable (Y) causa dos variables distintas (X y Z), generando una correlación (no causal) entre ambas.

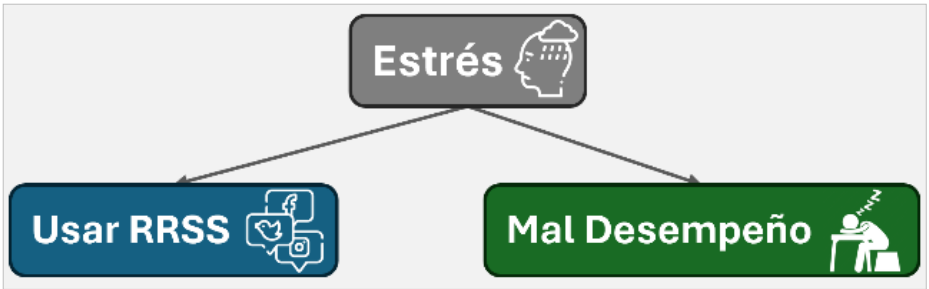
[20] *Idem*.
[21] J. PEARL, *op. cit.*

Para ilustrar la diferencia entre correlación y causalidad en el caso de que dos variables están correlacionadas debido a una causa común que genera las dos, consideremos el ejemplo de la correlación entre la instalación de luces navideñas en las calles españolas (X) y el aumento de peso de la población nacional (Z). Si observamos los datos, podríamos notar que el peso medio de la población española tiende a aumentar durante las semanas en las que se instalan luces navideñas en las ciudades. Aunque esta correlación se repite año tras año, es evidente que la instalación de luces navideñas no causa un aumento de masa corporal en los ciudadanos españoles, sino que la causa de este es una mayor ingesta calórica, derivada de las numerosas y copiosas comidas navideñas. En este caso, estaríamos ante una correlación entre dos variables explicada por una causa común, la navidad (Y), no ante una relación de causalidad, y, por tanto, a ningún jurista se le ocurriría la idea de considerar la instalación de luces navideñas como causa de un engordamiento generalizado.

Desde la perspectiva humana, gracias a nuestro conocimiento del contexto y de las circunstancias concurrentes, es fácil discernir que la causalidad no está presente en este caso. Entendemos que poner o no las luces de navidad no influye en que las personas engorden. Por ejemplo, si en una navidad no se instalasen tales adornos lumínicos por recortes presupuestarios, el incremento de peso seguiría produciéndose debido a las comidas copiosas que seguirían haciéndose durante el periodo navideño. Sin embargo, un sistema de IA, que basa su conocimiento del entorno en encontrar correlaciones en los datos con los que ha sido entrenado, y que carece de una observación directa del mundo, no tiene la capacidad para generalizar más allá del conocimiento presente en esos datos. Por lo tanto, si entrenáramos a un sistema de IA para predecir el aumento del peso corporal medio de la población española, es posible que el sistema correlacionase la instalación de luces con el aumento de peso corporal, prediciendo un aumento de peso o masa corporal durante las semanas en las que se instalan las luces navideñas (lo que estadísticamente sería cierto), a pesar de que estas dos variables no están causalmente relacionadas.

Más específicamente, en el ámbito laboral, podemos ejemplificar cómo observar una correlación que es causada por un factor común externo que no conocemos puede influir en la errónea identificación de una causa positiva. Así, supongamos que una empresa está investigando la relación entre el uso de redes sociales (RS) durante las horas de trabajo y el desempeño laboral de sus empleados.

**Figura 4** – Ejemplo de cómo el nivel de estrés es la causa común del nivel de uso de RS y del nivel de desempeño en el trabajo



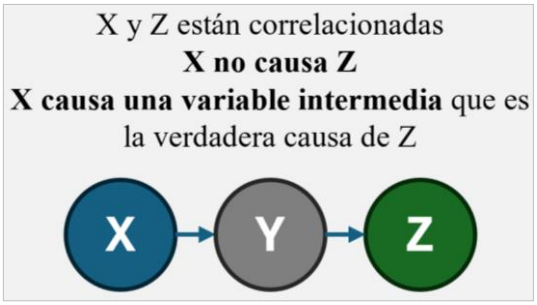

El sistema de IA concluye que hay una correlación negativa entre el uso de RS durante las horas de trabajo (X) y el desempeño laboral (Z), al detectar que los empleados que pasan más tiempo en las RS tienden a tener un desempeño laboral inferior. Sin embargo, el análisis inicial no tuvo en cuenta un factor importante: el nivel de estrés en el trabajo (riesgo laboral psicosocial) es la causa de ambas variables. Resulta que los empleados que experimentan niveles más altos de estrés en el trabajo tienden a buscar distracciones, como usar RS, durante sus horas de trabajo. Al mismo tiempo, el estrés laboral también puede afectar negativamente al desempeño laboral. Por lo tanto, la verdadera causa detrás de la correlación observada entre el uso de RS y el desempeño laboral no es el uso de RS en sí mismo, sino el nivel de estrés en el trabajo, actuando este como una variable causante de las otras dos, que influye tanto en el uso de RS como en el desempeño laboral, creando una correlación entre ambas, sin que exista una relación causal entre ellas.

En este ejemplo, el análisis inicial podría malinterpretar la relación entre el uso de RS y el desempeño laboral, debido a la falta de consideración del factor común externo (estrés en el trabajo). Si intervenimos para reducir el uso de RS, el desempeño no mejorará porque no es la causa directa. En cambio, si reducimos el estrés se reducirá el uso de RS a la vez que se mejora el desempeño laboral.

*2) Ambas variables están relacionadas a través de una variable intermedia*

En este caso las dos variables estén vinculadas mediante una variable intermedia llamada mediadora o *proxy*. De tal forma que una de las variables de interés causa una variable intermedia, y esa intermedia causa la otra variable de interés. Las dos variables (X y Z) estarán correlacionadas, aunque no mediante una relación causal.

**Figura 5** – Variables correlacionadas a través de una causa intermedia




Para entender más en profundidad este supuesto, vamos a ejemplificar un caso de encubrimiento de una discriminación laboral mediante la existencia de una variable intermedia o *proxy*. Imaginemos el uso de un sistema de IA diseñado para optimizar el proceso de selección de candidatos en una empresa. El sistema de IA ha sido entrenado con grandes conjuntos de datos, y ha detectado las correlaciones existentes en ellos (estén o no sesgadas), basándose en esas correlaciones para desarrollar la tarea de selección de candidatos para el puesto de trabajo. Si el *Dataset* de entrenamiento estuviese sesgado y reflejase estadísticamente una menor contratación de personas de determinados barrios, el sistema de IA se limitará a reproducir y a amplificar en sus predicciones ese sesgo.

El problema surge cuando el hecho de vivir en un determinado barrio es debido a bajos ingresos o a la pertenencia a una determinada etnia o religión. A ningún jurista se le ocultaría que no contratar a una persona por dichos motivos es una discriminación laboral que no tiene encaje en nuestro marco legal. Sin embargo, el problema surge cuando existe una variable proxy entre los bajos ingresos, la etnia o la religión y la contratación de esos candidatos, variable proxy que en este caso sería el Código Postal (CP) del candidato al puesto de trabajo y que, a pesar de que aparentemente tendría una naturaleza jurídicamente inocua, estaría encubriendo una auténtica discriminación laboral. Gráficamente, puede ser representado como sigue.

**Figura 6** – Ejemplo de cómo el CP (causa identificada), sirve de proxy del nivel de ingresos (causa del CP), lo cual indica una causa negativa si usamos el CP como variable de decisión


En conclusión, existen tres grandes grupos de correlaciones: las correlaciones espurias (en las que hay casualidad, pero no causalidad), las correlaciones causales (en las que hay causalidad, pero no casualidad) y las correlaciones influidas por factores externos no identificados (en las que hay causalidad, pero desconocimiento de cuál es la causa). El conocimiento de estos conceptos es necesario en el uso laboral de los sistemas de IA, pues en los mismos se está delegando paulatinamente el desarrollo de funciones empresariales que en muchas ocasiones exigen la concurrencia y suficiencia de causas legales, o la ausencia de causas discriminatorias. Todo lo cual puede complicarse por el efecto *black box* que entorpece la explicabilidad de las decisiones adoptadas o propuestas por la IA.

### 4.3.3. *Black box*. Explicabilidad

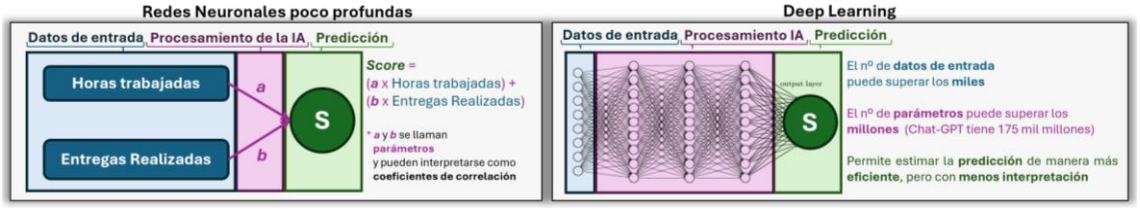
La identificación de relaciones causales a través de correlaciones se vuelve aún más desafiante en el actual contexto de los modelos de IA basados en DL, donde los modelos disponen de un gran número de parámetros, manifestando un funcionamiento complejo, y siendo extremadamente difícil conocer las correlaciones existentes en los datos[22]. Los modelos de DL son conocidos como modelos de caja negra (*black box*), ya que su funcionamiento interno es muy opaco, siendo difícil o imposible para un humano su comprensión y explicación. Mientras que en modelos simples podemos entender e interpretar fácilmente las correlaciones entre las variables, en el DL, los modelos están compuestos por múltiples capas de neuronas interconectadas, lo que dificulta la comprensión de cómo se toman las decisiones a partir de los datos de entrada[23]. Estos modelos hallan automáticamente correlaciones en los conjuntos de datos, aprendiendo patrones y características complejas de estos, lo que posibilita que sean

[22] I. GOODFELLOW, Y. BENGIO, A. COURVILLE, *op. cit.*

[23] C. PANIGUTTI, R. HAMON, I. HUPONT *ET AL.*, *The role of explainable AI in the context of the AI Act*, in VV.AA., *Proceedings of the 6th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2023)*, ACM, 2023.

altamente efectivos en tareas de clasificación y predicción, pero a costa de una limitada interpretabilidad y explicabilidad de sus decisiones. La diferente complejidad de la interpretabilidad de los modelos clásicos y el DL puede ser mejor entendida con la siguiente Figura.

**Figura 7** – Explicación de por qué es difícil incluso estimar las correlaciones en los modelos de DL




Los ejemplos utilizados en el epígrafe anterior son muy básicos y sencillos, pudiendo asemejarse al esquema de la izquierda de la Figura 7. Por el contrario, la realidad suele adecuarse más al esquema de la derecha en la que el sistema de IA basa su funcionamiento en miles de millones de parámetros procesados en innumerables capas de una red neuronal. Pero si comprender las correlaciones y causalidades de los ejemplos anteriores puede revestir cierta complejidad, lo cierto es que la comprensión de supuestos con billones de variables y parámetros lo es aún más. Es esta extrema dificultad o incluso imposibilidad lo que motiva a los científicos a denominar a estos modelos de IA modelos *black box* o de caja negra.

Más concretamente, en el ámbito del Derecho del Trabajo, la existencia y suficiencia de causa legal en la toma de decisiones empresariales es un requisito esencial en la validez de estas. Por ello, si la causa es exigida cuando la decisión es tomada por un ser humano (empresario o encargado), lo lógico es considerar que también debe ser exigida cuando la decisión empresarial es delegada en un sistema de IA (denominada "gestión algorítmica"). En otras ocasiones se podrían generar correlaciones que generasen decisiones empresariales basadas en causas discriminatorias difíciles de identificar. Sin embargo, las limitaciones técnicas que estos sistemas de IA manifiestan durante su funcionamiento, hacen difícil o incluso imposible constatar la existencia de posibles vicios relativos a la causa de la gestión algorítmica, lo que genera una incertidumbre sobre su legalidad y por ello una inseguridad jurídica difícil de paliar.

Para solventar este problema la normativa europea y nacional exigen que los sistemas de IA sean interpretables y explicables, y por consiguiente, transparentes. De esta forma, la transparencia se erige como un requisito esencial de una IA fiable y se hace efectiva en la práctica mediante la

información algorítmica contemplada en el art. 64.4.*d* del Texto Refundido del Estatuto de los Trabajadores (ET) aprobado por RDL 2/2015, de 23 de octubre. Esta información debe proporcionarse en aquellos casos en los que las facultades empresariales hayan sido delegadas en los sistemas de IA. En este supuesto la persona trabajadora es sometida a unas decisiones o a unas propuestas de decisiones generadas por una máquina que es opaca en su funcionamiento, lo que genera una indefensión en la persona destinataria de estas, que desconoce las causas y los procesos internos del sistema de IA.

Esta información, que deberá ser útil y comprensible, será aportada por el empresario a la representación legal y sindical de las personas trabajadoras (o en su defecto a la persona afectada), lo que les permitirá comprobar el cumplimiento de la normativa laboral aplicable, y en su caso, valorar la posible interposición de una denuncia en vía administrativa (ante la Inspección de Trabajo y Seguridad Social – ITSS) o en vía judicial (ante los juzgados y tribunales del orden social).

Por último, hay que indicar que existen medios y enfoques técnicos para facilitar la identificación de la causa y evitar la opacidad de los modelos de IA, a estas soluciones nos referiremos en el § 6.

## 5. Existencia de causalidad en la gestión algorítmica

Una vez que se han expuesto sucintamente los conceptos técnicos más relevantes de la materia, procede ahora efectuar la intersección con el Derecho del Trabajo español, lo que permitirá delimitar de forma visible el problema objeto de estudio. Este problema, al que denominaremos "dilema correlación-causalidad" se basa en el hecho de efectuar una delegación de funciones empresariales que requieren la concurrencia de causas legales (no de correlaciones) en sistemas de IA que basan su funcionamiento en correlaciones (no en causas). Para su análisis examinaremos las diferentes disposiciones de Derecho del Trabajo relativas a la concurrencia de causa legal y ausencia de causa discriminatoria en las decisiones empresariales. Por otra parte, al igual que en el ámbito técnico se diferencian tres supuestos en materia de correlación-causalidad, se ha considerado necesario crear por analogía dos nuevos conceptos, para facilitar un estudio más estructurado y clarificado del problema aquí examinado, la dimensión positiva y la dimensión negativa. Podemos entender los mismos como:

- la dimensión positiva orbita sobre la existencia de causa legal en la toma de decisiones por sistemas de IA. En el Derecho laboral el legislador ha querido que en muchas decisiones empresariales de

gestión de las personas trabajadoras tales como la modificación sustancial de condiciones de trabajo (MSCT), el régimen disciplinario o los despidos se exija como necesaria la concurrencia de una causa que sirva como justificación de la decisión empresarial. Aplicando los conceptos anteriormente expuestos podemos concluir que el Derecho laboral indica que si pasa X (por ejemplo, una causa económica, técnica, organizativa o de producción), entonces el empresario podrá decidir Y (por ejemplo, una MSCT o un despido por causas objetivas). Es decir, la norma establece la necesaria existencia de una relación causal entre la concurrencia de unas circunstancias previamente contempladas por la ley, y la posible adopción de una determinada decisión por parte del empresario. Es lógico considerar que si dicha relación causal es exigida cuando la decisión es tomada por un ser humano, también deberá ser exigida cuando la misma sea tomada por un sistema de IA. Por lo tanto, el estudio de la necesaria concurrencia de causa legal en la gestión algorítmica lo denominaremos "dimensión positiva" (del dilema correlación-causalidad);

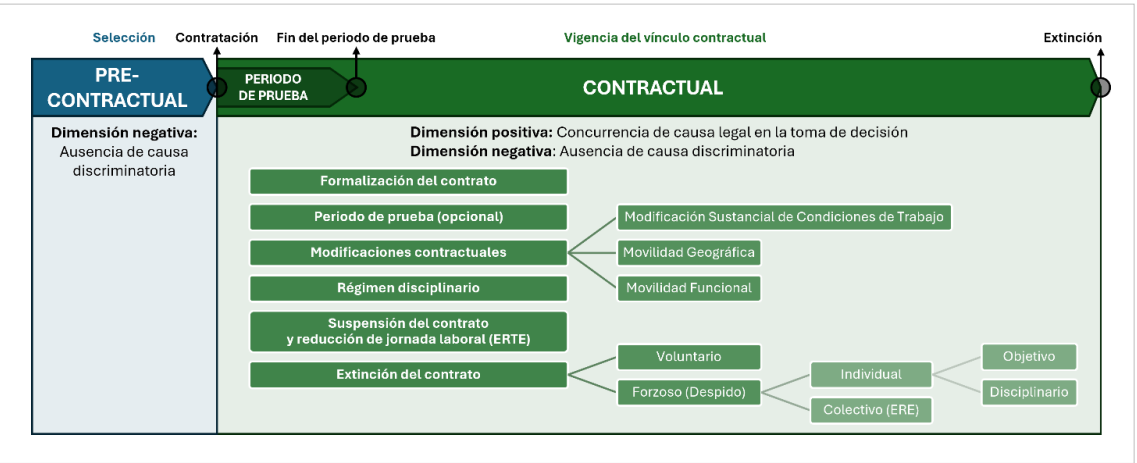
- la dimensión negativa gira en torno a la posible existencia de causas discriminatorias en las decisiones adoptadas por un sistema de IA. En el Derecho laboral el legislador también ha querido que las decisiones empresariales no estén motivadas por determinadas causas que generen discriminaciones directas o indirectas. Concretamente, el primero apartado del art. 17.1 ET establece una serie de causas que no podrán motivar las decisiones empresariales, y en caso de concurrir las mismas, la decisión empresarial se declarará nula; indicándose en su segundo apartado que también serán nulas las órdenes de discriminar. Por todo ello, debemos considerar lógico que, si la nulidad se extiende tanto a las decisiones discriminatorias como a las órdenes de discriminar, la nulidad sea también extensible a las decisiones discriminatorias de los sistemas de IA (la llamada discriminación algorítmica)[24]. Aplicando los conceptos técnicos anteriormente expuestos, podrá advertir el *ius laboralista* que a veces los sistemas de IA pueden generar decisiones discriminatorias basadas en causas difíciles de detectar e identificar. Al estudio de la inexistencia de causa discriminatorias en las decisiones adoptadas por sistemas de IA lo denominaremos "dimensión negativa" (del dilema correlación-causalidad).

Por último, para exponer de forma estructurada el estudio de ambas

[24] M. KEARNS, A. ROTH, *op. cit.*

dimensiones y su impacto en las relaciones laborales, se expone a continuación un esquema en el que se indican las diferentes decisiones empresariales afectadas por el dilema correlación-causalidad, diferenciando entre la fase precontractual y la fase contractual.

**Figura 8** – Esquema de elaboración propia en el que se refleja las diferentes decisiones empresariales que exigen la existencia y suficiencia de causas y que pueden afectar a las personas trabajadoras




## 5.1. Fase precontractual

En la fase precontractual, en la que todavía no se ha formalizado contrato laboral de trabajo alguno, el empleador con base en el derecho a la libertad de empresa de conformidad al art. 38 de la Constitución Española (CE), tiene la posibilidad de contratar o no a las personas trabajadoras, y cuando opte por ello, tiene libertad para elegir entre una u otra. No obstante, dicha contratación está sometida al principio de legalidad contemplado en el art. 53 CE, siendo uno de los límites legales a la libertad de selección y contratación de las personas trabajadoras el respeto al principio de no discriminación consagrado en el art. 14 CE.

Por tanto, la discriminación puede surgir durante el proceso selectivo, por ejemplo, rechazando sistemáticamente currículums de personas de un determinado sexo. En los casos en los que se materializan estas discriminaciones suelen surgir problemas probatorios que son suavizados mediante las previsiones normativas contempladas en los arts. 96.1, 179.2 y 181.2 de la Ley 36/2011, de 10 de octubre, reguladora de la jurisdicción social (LJS). En caso de falta de contratación de una persona por causas

discriminatorias la posible indemnización a la persona afectada se establecería de conformidad al art. 1106 del Código Civil.

Desde un punto de vista administrativo, existen dos conductas que responden a la posible discriminación en el acceso empleo, pudiendo generarse las mismas en dos momentos diferentes[25]:

1. la situación previa, cuando se soliciten datos de carácter personal en los procesos de selección o se establezcan condiciones, mediante la publicidad u otro medio que constituyan discriminación para el acceso al empleo. Dicha conducta podrá ser calificada como infracción administrativa muy grave en materia de empleo de conformidad al art. 16.1.*c* TRLISOS[26];
2. la formalización de la contratación bajo actos discriminatorios. Esta conducta podrá ser calificada como infracción administrativa muy grave en materia de relaciones laborales de conformidad al art. 8.12 TRLISOS.

Todas estas cuestiones deben ser examinadas desde la perspectiva de la dimensión positiva y negativa del dilema correlación-causalidad.

*Dimensión positiva*

En la fase precontractual el legislador no exige concurrencia de causa legal alguna que justifique su decisión de contratación, únicamente exige que no existan causas discriminatorias en el acceso al empleo. Esto supone que en esta fase solo se podrán generar problemas en la dimensión negativa.

*Dimensión negativa*

El derecho laboral permite al empresario realizar las pruebas de selección que considere convenientes (entrevistas, test psicotécnicos o pruebas de aptitud) para averiguar la capacidad profesional o aptitud de la persona candidata al puesto de trabajo, pero no puede indagar en datos que correspondan a su esfera íntima y personal. Así, se prohíben las indagaciones sobre la ideología (política, sindical) o creencias religiosas de los trabajadores, y sobre aspectos como la vida sexual, estado civil o datos familiares, no pudiendo, además, establecer criterios discriminatorios en el acceso al empleo por las circunstancias contempladas en el art. 17.1 ET.

[25] VV.AA., *Mémento Social 2023*, Francis Lefebvre, 2023.
[26] RDL 5/2000, de 4 de agosto, por el que se aprueba el texto refundido de la Ley sobre Infracciones y Sanciones en el Orden Social.

En la fase precontractual, la dimensión negativa se genera al automatizar o semiautomatizar las funciones de selección de personal mediante sistema de IA, que pueden:

a. inferir y predecir datos, en ocasiones sensibles (art. 9 RGPD), de las personas candidatas que no han sido aportados por las mismas. Por ejemplo, inferir el sexo de una persona por la letra manuscrita o el origen racial o étnico por razón del CP reflejado en el currículum[27];
b. hallar correlaciones entre datos (aparentemente inocuos), y basar la decisión en una causa discriminatoria difícil de detectar e identificar. Por ejemplo, rechazar sistemáticamente currículums en los que se refleje un determinado CP, o excluir sistemáticamente la contratación de mujeres al haber inferido su sexo de otra información de naturaleza aparentemente neutra.

En ambos casos, se podría estar incurriendo en las infracciones administrativas muy graves indicadas anteriormente, pero cuya detección por la ITSS sería muy difícil debido a la complejidad inherente a esta materia, que se conjugaría con las limitaciones técnicas de los sistemas de IA anteriormente expuestas.

Ejemplo de la dimensión negativa en la fase precontractual ha sido el uso por la empresa Amazon de una herramienta de filtrado de candidatos mediante un sistema de IA que generaba una discriminación encubierta de mujeres[28]. El sistema correlacionaba las carreras técnicas, y por ello la mayor probabilidad de contratación, con personas de sexo masculino (hecho que es estadísticamente mayoritario), por lo que las personas de sexo femenino obtenían una menor probabilidad de ser seleccionadas por el sistema. En este caso, el modelo de IA había sido entrenado con datos sesgados en los que solo había hombres con la formación adecuada, proponiendo la contratación de los candidatos por su sexo y no por su formación. Otro ejemplo es el caso del uso del programa de la empresa HireVue que priorizaba a los candidatos que hablaban despacio, siendo una característica que correlacionaba con los hombres, generando una discriminación laboral indirecta[29].

---

[27] A. TODOLÍ SIGNES, *Algoritmos productivos y extractivos. Cómo regular la digitalización para mejorar el empleo e incentivar la innovación*, Aranzadi, 2023, p. 63.

[28] THE GUARDIAN, *Amazon ditched AI recruiting tool that favored men for technical jobs*, en *www.theguardian.com/europe*, 11 octubre 2018.

[29] A. TODOLÍ SIGNES, *op. cit.*, p. 61.

## 5.2. Fase contractual

En la fase contractual cabe diferenciar las siguientes decisiones empresariales.

*a) Formalización del contrato de trabajo*

Respecto a la formalización del contrato, debemos hacer referencia a dos realidades estrechamente conectadas pero diferentes, cuya distinción es necesaria efectuar a los efectos que aquí nos atañen: la decisión de contratar y la formalización del contrato de trabajo. La selección o no de una persona para ser contratada es una decisión, siendo la formalización del contrato la materialización de dicha decisión, es decir, un contenido. Así, diferenciamos entre:

a. la elección o no de contratar a una persona es una decisión que queda incluida dentro de la fase precontractual anteriormente expuesta, y al ser una decisión podrá ser automatizada o semiautomatizada mediante la IA predictiva (*vid.* § 5.1);
b. el contrato de trabajo es un acuerdo, que puede recoger el contenido que materializa las condiciones del vínculo laboral entre empresario y persona trabajadora. Por tanto, no es una decisión, sino el fruto de una decisión (de contratación), que podrá ser automatizado o semiautomatizado mediante una IA generativa.

En nuestro ordenamiento jurídico la mayoría de los contratos de trabajo son causales, debiendo existir una coincidencia entre la modalidad contractual y la necesidad productiva presente en la empresa. El empresario tiene libertad para contratar, pero no libertad para elegir la modalidad contractual con la que se vincule con el trabajador[30]. En el presente epígrafe tampoco desarrollaremos esta cuestión al centrarse este artículo en la IA predictiva y no en la generativa.

*b) Periodo de prueba*

El periodo de prueba en nuestro ordenamiento jurídico está regulado en el art. 14 ET. Su finalidad consiste en que en ambas partes experimenten y se cercioren de que la relación laboral responde a la satisfacción de los intereses de cada uno. El periodo de prueba adquiere relevancia en materia

[30] En caso de no respetarse la causa del contrato, este se considera celebrado en fraude de ley, se transformará a indefinido a tiempo completo y podrá ser constitutiva dicho hecho como una infracción administrativa grave en materia de relaciones laborales.

de causalidad en lo que respecta a su extinción. En principio, la decisión extintiva por cualquier parte no exige justificación, pero es necesario que, si la misma es adoptada por la parte empresarial, esta no se fundamente, no esté motivada en una causa discriminatoria o viole derechos fundamentales[31].

*c) Modificaciones contractuales*

Durante la vigencia del vínculo contractual el empresario puede introducir modificaciones en algunos de los aspectos laborales que inicialmente fueron pactadas en el contrato de trabajo. Estas modificaciones pueden ser las siguientes.

*c.1) Movilidad funcional*

El objeto del contrato de trabajo está delimitado por la clasificación profesional del trabajador, pudiendo el empresario exigir a la persona trabajadora en principio cualquier función que este dentro de la misma. La movilidad funcional hace referencia a la exigencia de tareas fuera de esa previa clasificación que se ha hecho de la persona trabajadora. La movilidad funcional puede ser:

- movilidad funcional horizontal (dentro del grupo profesional). En materia causal debe indicarse que los convenios colectivos puedan en ocasiones condicionarla a "necesidades de servicio" o a "necesidades de la organización y dirección" y que se debe respetar el principio de no discriminación;
- movilidad funcional vertical (fuera del grupo profesional). En materia causal debe indicarse que para realizar la movilidad funcional vertical es necesario que concurra una causa técnica u organizativa que la justifique (art. 39.2 ET);
- movilidad funcional extraordinaria (cambio de funciones no previstas en el art. 39 ET). A efectos de exigencia de causa, cabe destacar que cuando hay pacto novatorio entre las partes la misma no será exigible, y en el resto de los casos habrá que estar a lo dispuesto en el convenio colectivo y en su defecto a la existencia de las causas contempladas en el art. 41 ET.

[31] *Cfr.* STC 38/1981, de 23 de noviembre; STC 94/1984, de 16 de octubre.

*c.2) Movilidad geográfica*

Consiste en el cambio del lugar donde se ejecuta la prestación de servicios, pudiendo ser calificado como un traslado o un desplazamiento. A efectos de concurrencia de causa, hay que destacar que en ambos casos se exige la concurrencia de causas económicas, técnicas, organizativas, de producción o cuando existan contrataciones referidas a la actividad empresarial. El trabajador podrá recurrir esta decisión empresarial ante la jurisdicción social de conformidad a lo dispuesto en los arts. 40 ET y 138 LJS.

*c.3) MSCT*

La MSCT[32] implica la modificación unilateral por el empresario de las condiciones laborales inicialmente pactadas entre este y la persona trabajadora. En materia causal el art. 41 ET exige la concurrencia de causas económicas, técnicas, organizativas o de producción, siendo competencia del orden social de la jurisdicción (*ex* art. 138 LJS) controlar la existencia de la causa alegada por el empresario y de la razonable adecuación entre la causa acreditada y la modificación acordada, y también la posible vulneración de los derechos fundamentales (como el principio de no discriminación). Cabe indicar que, desde un punto de vista causal, la sentencia judicial podrá declarar la MSCT como justificada (si concurre la causa), injustificada (si existe irregularidad causal) o nula (si hay lesión de derechos fundamentales, en lo que aquí atañe, vulneración del principio de no discriminación).

*d) Régimen disciplinario*

El poder disciplinario (*ex* art. 58 ET) es una consecuencia del reconocimiento del poder de dirección empresarial, que deberá respetar la graduación de faltas y sanciones que se establezcan en las disposiciones legales o en el convenio colectivo aplicable. Por consiguiente, se exige una doble tipicidad, la de las infracciones y la de las sanciones. En caso de que se ha sancionado a una persona trabajadora, dicha decisión empresarial podrá ser revisada por el orden social de la jurisdicción (art. 58.2 ET). La sentencia judicial podrá confirmar la sanción o revocarla (total o parcialmente), dependiendo de si se ajusta o no a las disposiciones

[32] Deben tenerse en cuenta también las previsiones del art. 82.3 ET en materia de descuelgue del convenio colectivo.

aplicables, o declararla nula si se comprueba que la sanción atiende a criterios discriminatorios o viola derechos fundamentales de la persona trabajadora.

Aunque en este caso no existe una remisión al concepto de causa, sí que es necesario la previa tipificación de las infracciones y sanciones para poder ejercer la facultad disciplinaria.

*e) Suspensión de contrato o reducción de jornada (ERTE)*

Existen una serie de supuestos que pueden motivar las suspensiones de los contratos de trabajo. Uno de esos supuestos es la concurrencia de causas económicas, técnicas, organizativas, de producción o de fuerza mayor, además de la posibilidad de adoptar dicha suspensión como medida disciplinaria. El resto de los supuestos son situaciones objetivas y reales con un reducido margen de interpretabilidad, que, en principio, no generarían problemas si se delega a un sistema de IA.

Durante la tramitación administrativa del ERTE la Autoridad Laboral, recabará informe de la ITSS (art. 22, RD 1483/2012), el cual deberá pronunciarse sobre la inexistencia de criterios discriminatorias en la designación de las personas afectadas por el mismo.

Contra las decisiones empresariales de suspensión de contratos o reducción de jornada podrán reclamar las personas trabajadoras ante la jurisdicción social, que podrá declarar la medida justificada o injustificada.

*f) Extinción del contrato*

Supone la rescisión del vínculo contractual. Tradicionalmente se clasifican en voluntarias (por voluntad de la persona trabajadora), forzosas (por voluntad empresarial) y por otras causas (jubilaciones, fallecimientos, etc.), siendo las dos primeras las que son adoptadas por una voluntad humana, y por ello, su delegación en un sistema de IA puede plantear problemas. Así, diferenciamos entre las siguientes.

*f.1) Voluntaria*

Cuando la extinción del vínculo contractual se realiza a instancia de la persona trabajadora tiene carácter voluntario. Como se trata de una extinción no basada en una decisión empresarial, sino en una decisión de desistimiento de la persona trabajadora, quedará claramente fuera del objeto de este estudio, y por ello, no será objeto de desarrollo.

*f.2) Forzosa*

Es aquella que es efectuada a instancia de del empresario, denominándose despido. Dentro de las extinciones contractuales por voluntad empresarial, cabe distinguir entre las siguientes.

*f.2.1) Individual (despido por causas objetivas o disciplinarias)*

El despido individual puede estar motivado por causas objetivas o causas disciplinarias:

- despido por causas objetivas (*ex* art. 53 ET). El empresario tiene la obligación de indicar la causa que lo motiva, y los hechos acontecidos (para evitar una indefensión de la persona despedida). El despido podrá ser impugnado en vía judicial por la persona afectada, pudiendo el órgano judicial declarar tal decisión justificada (cuando se acredite la concurrencia de causa), improcedente (cuando no se acredite la concurrencia de causa) o nula (cuando se acredite que estuvo motivada por una causa discriminatoria prohibida por la CE o una por violación de los derechos fundamentales y libertades públicas);
- despido por causas disciplinarias (*ex* art. 54 ET). El empresario tiene la obligación de indicar la causa que lo motiva, debiéndose incardinar en alguna de las descritas en el art. 56.2 ET. Ahora bien, no rige el principio de tipicidad legal con la misma intensidad que en el ámbito del derecho sancionador del Estado (STC 69/1983, de 26 de julio); ello significa que estas causas legales incluyen un amplio espectro de supuestos concretos, teniendo el convenio colectivo un amplio margen para concretar las causas legales (STS de 17 octubre 2023).

El empresario de conformidad al art. 55.1 ET deberá notificar el despido por escrito, indicando los hechos que lo motivan de forma clara y concisa (no siendo suficiente una mera referencia genérica a determinados hechos), así como la fecha en que tendrá efectos. La información que ha de contener la carta de despido tiene por finalidad que evitar la indefensión de la persona afectada por el despido, que podrá ser impugnado ante la jurisdicción social de conformidad al art. 103 ss. LJS. La resolución judicial podrá calificar el despido como procedente (cuando se acredite la concurrencia de causa), improcedente (cuando no se acredite la concurrencia de causa) o nulo (cuando se acredite que estuvo motivada por una causa discriminatoria prohibida por la CE o por una violación de los derechos fundamentales y libertades públicas).

*f.2.2) Colectiva (ERE)*

Cuando el empresario adopta una decisión extintiva sobre un número de personas trabajadoras que supera los umbrales indicados en el art. 51.1 ET, el despido tendrá la calificación de colectivo, debiendo concurrir causas económicas, técnicas, organizativa, de producción, o de fuerza mayor. Adicionalmente, en materia causal debe indicarse que:

- durante la tramitación del procedimiento de ERE ante la Autoridad Laboral, la ITSS deberá emitir informe en el que se pronunciara sobre la suficiencia de la causa, así como que no ha habido discriminación alguna en la designación de las personas afectadas por dicho despido;
- la decisión empresarial podrá ser impugnada en vía judicial, y posteriormente calificada como: ajustada a derecho (cuando se acredite la existencia de la causa esgrimida), no ajustada a derecho (en caso contrario), o nula (entre otras razones, cuando se haya adoptado con vulneración de derechos fundamentales y libertades públicas), de conformidad al art.134 LJS.

*Dimensión positiva*

Las diferentes decisiones empresariales anteriormente expuestas exigen la concurrencia de causa legal, cuya existencia y suficiencia deberá ser justificada por el empresario, y posteriormente controlada por el orden social de la jurisdicción. Por ejemplo, en la MSCT el empresario «deberá aportar prueba de esa ligazón entre las causas aducidas, las medidas adoptadas y los efectos pretendidos», debiendo respetar en todo momento el principio de proporcionalidad[33], lo que adquiere especial relevancia debido a la aparente laxitud con la que está redactado el art. 41 ET; o en el caso de los despidos individuales o colectivos. La prueba puede ser realmente difícil de aportar como consecuencia de las características técnicas anteriormente expuestas que manifiestan los sistemas de IA.

Además, suele ser habitual también la valoración por la ITSS de la concurrencia de causa legal a través de informes solicitados por la Autoridad Laboral o los juzgados y tribunales del orden social de la jurisdicción, como en el caso de la tramitación de los ERE, o de los supuestos contemplados en el art. 138.3 LJS (MSCT, movilidad geográfica,

[33] ARANZADI, *Modificaciones del contrato de trabajo. Modificación sustancial de condiciones de trabajo*, DOC 2003\136, 2023.

suspensión de contratos, etc.). Más aún, en el caso de los despidos nótese que pueden ser calificados desde un punto de vista causal por el orden social de la jurisdicción como: procedente (cuando se acredita la misma), improcedente (cuando no se acredita) o nula (cuando se produce una vulneración de libertades públicas y derechos fundamentales como el principio de no discriminación). En los dos primeros casos (calificación procedente e improcedente) estaríamos ante un supuesto de dimensión positiva, es decir, de existencia y suficiencia de la causa, y en el tercer caso (calificación nula) ante la dimensión negativa, es decir, de inexistencia de causa discriminatoria. Por todo ello, puede inferirse claramente que la justificación causal adquiere una especial importancia cuando estas funciones empresariales son delegadas en sistemas de IA.

Nuevamente, el problema de la valoración de la existencia y suficiencia de causa surge cuando se automatizan o semiautomatizan estas decisiones empresariales, pues el sistema de IA no tiene capacidad para identificarlas, y en caso de que lo hiciese, surgen muchas dificultades para su constatación.

Todo ello genera una inseguridad jurídica que se extiende no solo a los derechos e intereses de las personas trabajadoras, sino también a los de los empresarios, y afecta frontalmente a la eficacia y eficiencia del poder judicial y de la ITSS, que en muchas ocasiones son desconocedores de las limitaciones técnicas de los sistemas de IA.

Ejemplos de la dimensión positiva de la fase contractual son:

- el caso de la empresa americana Xsolla, dedicada a servicios de pago en videojuegos, que efectuó un despido colectivo que afecto al 30% de la plantilla debido a una recomendación de un sistema de IA[34];
- el caso de la empresa Amazon que también ha utilizado sistemas de IA para automatizar despidos de sus personas trabajadoras sin intervención humana debido a bajos índices de productividad[35].

En el primer caso, la decisión empresarial se adoptó atendiendo únicamente a la recomendación del sistema de IA, obviando la opinión jurídica fundamentada del departamento de recursos humanos de la mercantil, y en el segundo se ejecutó de forma automatizada. En ambos casos, la justificación causal sería de difícil prueba y, por ende, de improbable encuadre dentro de nuestro marco legal.

---

[34] *Cfr.* M. ECHARRI, *150 despidos en un segundo: así funcionan los algoritmos que deciden a quién echar del trabajo*, en *elpais.com*, 10 octubre 2021; AIAAIC, *Xsolla uses secret monitoring system to fire employees*, en *www.aiaaic.org*, enero 2022.

[35] A. TODOLÍ SIGNES, *op. cit.*, pp. 18-19 y 45.

*Dimensión negativa*

La dimensión negativa también está presente en la fase contractual de toda relación de trabajo, pues en todo momento es necesario que las decisiones empresariales no estén motivadas por criterios discriminatorios, debiendo por ejemplo la ITSS indicar en el informe de ERTE si se han constatado motivos discriminatorios en la designación de las personas trabajadoras afectadas por el mismo. Nuevamente, es necesario indicar, que el sistema de IA fundamenta su funcionamiento en correlaciones, y en ocasiones pueden basar el mismo en causas discriminatorias encubiertas.

Un ejemplo de dimensión negativa podría ser las medidas disciplinarias adoptadas por el sistema de IA de forma semiautomatizada o automatizada como consecuencia de un comportamiento sesgado derivado de las puntuaciones de los clientes, como ha sido el caso de la plataforma digital Uber[36].

Todos los problemas derivados de la dimensión positiva y negativa del dilema de correlación-causalidad nos lleva a presentar algunas soluciones que eliminen o al menos reduzcan los riesgos derivados de su existencia.

## 6. Propuesta de soluciones

Al tratarse esta materia de una intersección entre los campos técnico y jurídico, se hace necesario proponer medidas relativas a cada uno de ellos, esto es, soluciones técnicas y soluciones jurídicas.

### 6.1. Soluciones técnicas

Para abordar los desafíos expuestos en el presente documento existen diferentes instrumentos y enfoques técnicos que desempeñan un papel crucial en mitigar la opacidad y el sesgo de los sistemas de IA en el ámbito laboral. Estas herramientas facilitan una mayor transparencia y comprensión sobre cómo se toman las decisiones, permitiendo una mejor supervisión humana, así como eliminar o al menos reducir la incertidumbre sobre la inexistencia, existencia y suficiencia de causa legal en la gestión algorítmica. Entre otras, podemos hacer referencia a las siguientes:

- justicia algorítmica (*Algorithmic Fairness*). Dentro de la dimensión

[36] A. RONSEBLAT, S. BAROCAS, K. LEVY, T. HWANG, *Discrimination Tastes. Customer Ratings as Vehicles for Bias*, Data & Society, 2016.

negativa, la justicia algorítmica[37] es un campo de estudio que se centra en garantizar que los algoritmos y los sistemas de IA sean justos y no perpetúen los sesgos existentes en los datos de entrenamiento, centrándose también en desarrollar algoritmos que aprendan sin discriminación;

- sistemas automatizados de ayuda a la toma de decisiones (*Automated Decision Support Systems* – ADSS). Son sistemas informáticos que apoyan el proceso de toma de decisiones, usando algoritmos y técnicas de IA para analizar grandes volúmenes de datos y generar recomendaciones precisas y relevantes, pero teniendo en cuenta una colaboración efectiva humano-máquina. En estos sistemas, la decisión final siempre depende del controlador humano, añadiendo el raciocinio humano y el entendimiento del contexto, aspectos que los sistemas de IA todavía no pueden reproducir completamente[38];
- inferencia causal. Es una técnica que cuantifica las relaciones de causa y efecto por encima de correlaciones, resulta esencial para entender la influencia de estas relaciones en decisiones automatizadas. Sin embargo, su aplicación (limitada) requiere un conocimiento profundo del dominio y habilidad para la interpretación de los resultados, ya que puede ser desafiante para aquellos sin formación en estadística o inferencia causal[39];
- la IA interpretable y la IA explicable (*eXplainable Artificial Intelligence* – XAI). Para hacer frente a la opacidad de los modelos de IA y facilitar su explicabilidad, desde el punto de vista técnico, existen dos grandes enfoques:
  a. la IA interpretable tiene por objetivo crear modelos de IA que sean diseñados desde sus inicios con un proceso interno de toma de decisiones que sea entendible por los seres humanos. Esto se logra, por ejemplo, utilizando modelos específicos o diseñando el modelo para procesar la información de una forma que sea interpretable. El inconveniente de este enfoque es que los modelos de IA interpretables suelen ser menos efectivos en el desarrollo de sus tareas;
  b. la XAI[40] tiene por objetivo hacer que los modelos de IA de caja negra sean más comprensibles, para ello permite el uso de estos

[37] S. BAROCAS, M. HARDT, A. NARAYANAN, *Fairness and Machine Learning. Limitations and Opportunities*, MIT Press, 2023.
[38] L. FLORIDI, J. COWLS, T.C. KING, M. TADDEO, *How to Design AI for Social Good: Seven Essential Factors*, en *Science and Engineering Ethics*, 2020, vol. 26, n. 3.
[39] J. PEARL, *op. cit.*
[40] C. PANIGUTTI, R. HAMON, I. HUPONT *ET AL.*, *op. cit.*

> modelos y se centra en proporcionar explicaciones sobre su proceso de toma de decisiones de manera fácilmente comprensible para las personas, sin imponer restricciones al propio modelo. Los métodos XAI pueden proporcionar explicaciones globales (cómo toman decisiones a nivel general) o locales (cómo toman decisiones para un caso específico) sobre el funcionamiento del sistema de IA.

Sin embargo, suele tener importantes limitaciones debido a la poca fiabilidad y solidez de las explicaciones, ya que es difícil saber qué variables han influido en la decisión. Además, las explicaciones son siempre aproximaciones imperfectas a los procesos internos de toma de decisiones de los modelos *black box*, no habiendo actualmente una forma clara ni consensuada de evaluarlas.

En conclusión, la IA interpretable es un enfoque científico basado en crear modelos de IA transparentes por diseño, simplificando el proceso interno de toma de decisiones del modelo de IA, mientras que la XAI se enfoca en interpretar modelos opacos (*black box*), no buscando simplificar el modelo, sino dar una explicación clara y comprensible de sus procesos internos.

## 6.2. Soluciones jurídicas

Asimismo, presentamos las siguientes soluciones de índole jurídica.

- Permitir únicamente la automatización de decisiones no causales, siendo posible la semiautomatización de las causales. Es importante recordar que los sistemas de IA pueden automatizar tareas (adoptando decisiones) o semiautomatizar tareas (ofreciendo una asistencia a la adopción de dichas decisiones). Debido a que los sistemas de IA basan su funcionamiento en correlaciones (y no en causas), conjugado con la dificultad de interpretar los procesos internos de los modelos *black box*, se hace lógico considerar que en ningún caso podría automatizarse decisiones empresariales que exijan concurrencia de causa, pudiendo, sin embargo, ser utilizados como una asistente al controlador humano, siendo este quien debe cerciorarse de la existencia y suficiencia de causa legal y ausencia de causa discriminatoria en la toma de las decisiones.
- Alfabetización en materia de IA (*ex* art. 3.56, Reglamento IA). Otra de las necesarias soluciones es la alfabetización en materia de IA de las personas trabajadoras, de los empresarios y de los funcionarios del poder judicial y de la ITSS. Respecto a estos últimos, se ha hecho referencia a diferentes supuestos en los que competencialmente los funcionarios deben

pronunciarse sobre la dimensión positiva y negativa del dilema correlación-causalidad, siendo necesario para ello unos conocimientos técnicos y jurídicos adecuados, ya que en caso contrario se podrían producir riesgos para los derechos e intereses legalmente reconocidos. Esta capacitación técnica en nuevas tecnologías ya está prevista en el eje estratégico 5 de la Estrategia Nacional de Inteligencia Artificial.

- Reforma legislativa. Por último, a pesar de que el en el presente estudio se ha querido presentar soluciones concretas, útiles y lógicas a los problemas planteados, también es esencial poner de relieve que en el Derecho del Trabajo la solución más prudente y efectiva a largo plazo siempre es el entendimiento entre los interlocutores sociales, verdaderos protagonistas de toda la construcción *ius laboralista*, principalmente cuando la misma se materializa en una norma de rango legal. Por tanto, se quiere resaltar la importancia de una futura regulación legal de las cuestiones aquí planteadas, que tenga por objeto dar una respuesta a los problemas que acaban de referirse.

## 7. Conclusiones

El objetivo principal de este trabajo ha sido exponer sintéticamente el problema inherente a la delegación de funciones empresariales que requieren la concurrencia de causas legales (no de correlaciones) en sistemas de IA que basan su funcionamiento en correlaciones (no en causas), lo que hemos denominado el "dilema correlación-causalidad". Esta praxis empresarial se está extendiendo progresivamente, pudiendo generar inseguridad jurídica en cuanto a la validez de tales decisiones.

Debido a la complejidad inherente a esta nueva realidad laboral, se ha considerado la necesidad de tener en cuenta, no solo conocimientos jurídico-laborales, sino también conocimientos propios de las ramas técnicas, buscando la concurrencia interdisciplinar de ambos sectores profesionales. Fruto de esta colaboración se han creado dos nuevos conceptos jurídicos (la dimensión positiva y la dimensión negativa de la causalidad), cuyo uso podrá facilitar futuros estudios de este problema. Más concretamente, la dimensión positiva (por inexistencia o insuficiencia de la causa legal exigida por el legislador) y la negativa (por generación de una causa encubierta que genera una discriminación laboral indirecta).

En conclusión, se ha querido centrar el foco de atención en los problemas legales vinculados con la causalidad en la toma de decisiones empresariales delegadas en sistemas de IA, así como poner de manifiesto el imprescindible conocimiento y uso de conceptos técnicos que hacen

necesaria la debida colaboración con investigadores de la vertiente técnica.

## 8. Bibliografía

AIAAIC (2022), *Xsolla uses secret monitoring system to fire employees*, en *www.aiaaic.org*, enero

ARANZADI (2023), *Modificaciones del contrato de trabajo. Modificación sustancial de condiciones de trabajo*, DOC 2003\136

BAROCAS S., HARDT M., NARAYANAN A. (2023), *Fairness and Machine Learning. Limitations and Opportunities*, MIT Press

BAROCAS S., SELBST A.D. (2016), *Big Data's Disparate Impact*, en *California Law Review*, vol. 104, n. 3, pp. 671-732

BARRY E. (2021), *Uber Drivers Say a 'Racist' Algorithm Is Putting Them Out of Work*, en *time.com*, 12 octubre

CHRISTENKO A., JANKAUSKAITĖ V., PALIOKAITĖ A., VAN DEN BROEK E.L., REINHOLD K., JÄRVIS M. (2022), *Artificial intelligence for worker management: an overview. Report*, EU-OSHA

DELIPETREV B., TSINARAKI C., KOSTIĆ U. (2020), *AI Watch. Historical Evolution of Artificial Intelligence. Analysis of the three main paradigm shifts in AI*, JRC Technical Report

ECHARRI M. (2021), *150 despidos en un segundo: así funcionan los algoritmos que deciden a quién echar del trabajo*, en *elpais.com*, 10 octubre

FISKE S.T. (1998), *Stereotyping, Prejudice, and Discrimination*, en D.T. GILBERT, S.T. FISKE, G. LINDZEY (eds.), *The Handbook of Social Psychology. Volume One*, McGraw-Hill

FLORIDI L., COWLS J., KING T.C., TADDEO M. (2020), *How to Design AI for Social Good: Seven Essential Factors*, en *Science and Engineering Ethics*, vol. 26, n. 3, pp. 1771-1796

GOERLICH PESET J.M. (dir.) (2023), *Derecho del trabajo*, Tirant lo Blanch

GOODFELLOW I., BENGIO Y., COURVILLE A. (2016), *Deep Learning*, MIT Press

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE (2019), *A Definition of AI: Main Capabilities and Scientific Disciplines*

KEANE J. (2021), *Deliveroo Rating Algorithm Was Unfair To Riders, Italian Court Rules*, en *www.forbes.com*, 5 enero

KEARNS M., ROTH A. (2019), *The Ethical Algorithm. The Science of Socially Aware Algorithm Design*, Oxford University Press

LÓPEZ DE MÁNTARAS BADIA R., MESEGUER GONZÁLEZ P. (2017), *Inteligencia artificial*, Catarata

OECD (2024), *Explanatory memorandum on the updated OECD definition of an AI system*, OECD Artificial Intelligence Paper, n. 8

OLIVER N. (2022), *Artificial intelligence for social good: the way forward*, en EUROPEAN COMMISSION, *Science, research and innovation performance of the EU 2022. Building a sustainable future in uncertain times*

OLIVER N. (2019), *Governance in the era of data-driven decision-making algorithms*, en A. GONZÁLEZ, M. JANSEN (eds.), *Women Shaping Global Economic Governance*, CEPR Press

PANIGUTTI C., HAMON R., HUPONT I. *ET AL.* (2023), *The role of explainable AI in the context of the AI Act*, in VV.AA., *Proceedings of the 6th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2023)*, ACM

PEARL J. (2009), *Causality. Models, Reasoning, and Inference*, Cambridge University Press

PETERS J., D. JANZING, B. SCHÖLKOPF, *Elements of Causal Inference. Foundations and Learning Algorithms*, MIT Press, 2017

RODRÍGUEZ FERNÁNDEZ M.L. (2024), *Inteligencia artificial, género y trabajo*, en *Temas Laborales*, n. 171, pp. 11-39

RONSEBLAT A., BAROCAS S., LEVY K., HWANG T. (2016), *Discrimination Tastes. Customer Ratings as Vehicles for Bias*, Data & Society

THE GUARDIAN (2018), *Amazon ditched AI recruiting tool that favored men for technical jobs*, en *www.theguardian.com/europe*, 11 octubre

TODOLÍ SIGNES A. (2023), *Algoritmos productivos y extractivos. Cómo regular la digitalización para mejorar el empleo e incentivar la innovación*, Aranzadi

TOLAN S., PESOLE A., MARTÍNEZ-PLUMED F., FERNÁNDEZ-MACÍAS E., HERNÁNDEZ-ORALLO J., GÓMEZ E. (2020), *Measuring the Occupational Impact of AI: Tasks, Cognitive Abilities and AI Benchmark*, European Commission

VIGEN T. (2015), *Spurious correlations. Correlation is not causation*, en *tylervigen.com*, 14 mayo

VV.AA. (2023), *Mémento Social 2023*, Francis Lefebvre

WILLSON M. (2017), *Algorithms (and the) everyday*, en *Information, Communication & Society*, vol. 20, n. 1, pp. 137-150

# Red Internacional de ADAPT

**ADAPT** es una Asociación italiana sin ánimo de lucro fundada por Marco Biagi en el año 2000 para promover, desde una perspectiva internacional y comparada, estudios e investigaciones en el campo del derecho del trabajo y las relaciones laborales con el fin de fomentar una nueva forma de "hacer universidad", construyendo relaciones estables e intercambios entre centros de enseñanza superior, asociaciones civiles, fundaciones, instituciones, sindicatos y empresas. En colaboración con el DEAL – Centro de Estudios Internacionales y Comparados del Departamento de Economía Marco Biagi (Universidad de Módena y Reggio Emilia, Italia), ADAPT ha promovido la institución de una Escuela de Alta Formación en Relaciones Laborales y de Trabajo, hoy acreditada a nivel internacional como centro de excelencia para la investigación, el estudio y la formación en el área de las relaciones laborales y de trabajo. Informaciones adicionales en el sitio *www.adapt.it*.

Para más informaciones sobre la Revista Electrónica y para presentar un artículo, envíe un correo a *redaccion@adaptinternacional.it*.

INSTITUTO DE INVESTIGACIONES JURÍDICAS
unam
ADAPT
www.adapt.it