Facultad de Ciencias Económicas y Empresariales
ICADE

EL USO DE LA INTELIGENCIA ARTIFICIAL EN LA CONCESIÓN DE PRÉSTAMOS BANCARIOS: PROBLEMAS ÉTICOS

Autor: Blanca López-Tello Martínez
Director: María Reyes Calderón Cuadrado

Madrid | Marzo 2025

Resumen: El uso de la Inteligencia Artificial por parte de las entidades financieras para la optimización de la concesión de préstamos es ya una necesidad ineludible y lo será en mayor medida en el futuro. Esta utilización puede conducir además de a una asignación eficiente de los recursos, a la apertura de los mercados financieros sin perjudicar la evaluación del riesgo crediticio. Ahora bien, la Inteligencia Artificial no es un simple programa informático, pues implica procesos propios de la mente. Así, el foco no solo debe ponerse en la ética de quien programa sino también en la forma en la que la IA desarrolla y aprende, pues los caminos que esta tome pueden dar lugar a resultados antiéticos. Este TFG tiene por objeto explorar los conflictos éticos que existen en el uso de la IA en la concesión de préstamos bancarios; e investigar las potenciales medidas para que ciertos problemas éticos no se produzcan y, en su caso, se mitiguen.

Palabras clave: Ética, Inteligencia Artificial, *machine learning*, concesión de préstamos, sesgo algorítmico, explicabilidad, opacidad, protección de datos.

Abstract: The use of Artificial Intelligence by financial institutions to optimise the granting of loans is already an unavoidable necessity and will be even more so in the future. This use can lead not only to an efficient allocation of resources, but also to the opening up of financial markets without impairing credit risk assessment. However, Artificial Intelligence is not a simple computer programme, as it involves processes of the mind. Thus, the focus should be not only on the ethics of the programmer but also on the way in which AI develops and learns, as the paths it takes can lead to unethical results. The aim of this dissertation is to explore the ethical conflicts that exist between the use of AI in bank lending; and to investigate potential measures to prevent certain ethical problems from occurring and, where appropriate, to mitigate them.

Key words: Ethics, Artificial Intelligence, machine learning, loan granting, algorithmic bias, explainability, opacity, data protection.

ÍNDICE

INTRODUCCIÓN AL TRABAJO

Para abordar las diferentes cuestiones que se pleantearán a lo largo del trabajo será necesario, en primer lugar y antes de nada, comprender aquello en lo que nos vamos a enfocar, esto es, la Inteligencia Artificial (de ahora en adelante, IA). En este sentido, diversas son las definiciones que podemos encontrar en relación con este concepto. Por un lado, el Parlamento Europeo define la IA como "la habilidad de una máquina de presentar las mismas capacidades que los seres humanos, como el razonamiento, el aprendizaje, la creatividad y la capacidad de planear" (Parlamento Europeo, 2020). En esta misma línea, el Consejo de la Unión Europea la define como "el uso de tecnología digital para crear sistemas que puedan realizar tareas que habitualmente requieren una intervención humana inteligente" (Consejo de la Unión Europea, 2025).

Estas definiciones ponen de manifiesto el carácter transformador de la IA y su creciente integración en sectores clave de la sociedad, como pueden ser el derecho, la educación o la banca (Stahl, 2022). Sin embargo, como toda tecnología disruptiva, la IA conlleva tanto beneficios como riesgos. Así, pese a que su capacidad para procesar grandes volúmenes de datos ha traído consigo multitud de avances significativos, la IA también ha suscitado debates con respecto a sus implicaciones éticas, convirtiéndose en una moneda de dos caras que debe ser desarrollada en un entorno de supervisión.

En este contexto, el presente trabajo se sitúa en el debate sobre el empleo de la IA en la concesión de préstamos bancarios y los desafíos éticos que esto puede conllevar. En efecto, la IA ha revolucionado la evaluación del riesgo crediticio al permitir análisis más precisos y eficientes, siendo la mayor representación de esto los modelos que hacen uso de aprendizaje automático o *machine learning*, una técnica central dentro de la IA que implica una capacidad de aprender a partir de los datos sin necesidad de una programación explícita. Este aprendizaje automático puede clasificarse en dos categorías: aprendizaje supervisado (el modelo recibe datos de entrada con un resultado esperado y debe aprender a establecer una relación entre ambos) y el aprendizaje no supervisado (el modelo recibe datos de entrada sin resultados y debe descubrir patrones en los datos). Además, existe una categoría que

cruza tanto el aprendizaje supervisado como el no supervisado, conocida como aprendizaje profundo (*deep learning*) (Aziz & Dowling, 2019).

Aunque el uso de estos sistemas ha conllevado muchos avances y mejoras, su implementación también ha generado preocupaciones. Así, el objetivo de este trabajo es analizar los riesgos éticos asociados al uso de IA en la evaluación crediticia y explorar las estrategias disponibles para mitigarlos. En este sentido, se llevará a cabo una revisión exhaustiva de la literatura académica sobre el uso de la IA en la concesión de préstamos bancarios. Este análisis incluirá la normativa vigente, artículos académicos, libros, informes de instituciones europeas que servirán para comprender e indagar en las distintas aplicaciones de la IA en la concesión de préstamos bancarios. A partir de la revisión de estas fuentes se realizará una evaluación crítica de los principales riesgos éticos planteados por el uso de esta inteligencia. Finalmente, se elaborará un análisis detallado de las herramientas y recursos disponibles a efectos de proponer estrategias para bien evitar o atenuar los problemas éticos identificados en el uso de IA.

Este estudio anticipa que la IA puede mejorar la evaluación del riesgo crediticio, reduciendo las desigualdades y mejorando el acceso al crédito, pero también puede reforzar prácticas discriminatorias en la toma de decisiones. Así, su implementación debe controlarse para evitar sesgos, garantizar la explicabilidad y proteger la privacidad. En este sentido, cabe mencionar que el análisis realizado en el presente trabajo se centra en el contexto europeo, pudiendo no ser aplicable a otras jurisdicciones, además de estar sujeto a cambios debido a la rápida evolución de la IA.

CAPÍTULO I. MARCO GENERAL DE LA IA EN EL SECTOR BANCARIO

1. LA APLICACIÓN DE IA EN LA CONCESIÓN DE PRÉSTAMOS BANCARIOS

En primer lugar, antes de adentrarnos en los desafíos éticos que puede plantear el uso de la IA en la concesión de préstamos bancarios, deben abordarse también los usos y ventajas que estos sistemas pueden aportar al crédito pues, en efecto, una cara de la moneda no puede comprenderse sin su opuesta. Así, el objeto del presente apartado es enunciar y describir las distintas aplicaciones de la IA en la concesión de préstamos.

1.1. Precisión en la evaluación del riesgo crediticio

El crédito es sin duda la actividad comercial fundamental para las entidades bancarias, ocupándose estas de evaluar y gestionar el riesgo asociado a la concesión de préstamos. En este sentido, los expertos banqueros consideran que el mayor peligro al que se enfrenta el sector bancario global es la morosidad en los préstamos, representando la principal fuente de presión en las operaciones bancarias (Ridzuan et al., 2024). Dada esta realidad, el primer y principal uso de la IA en el sector bancario es el de la evaluación del riesgo crediticio.

Para medir la solvencia de los clientes, los bancos hacen uso de una gran recopilación de datos de carácter transaccional, de análisis estadísticos, árboles de decisión y regresión para determinar si el consumidor ostenta la capacidad para devolver un crédito (Federación Bancaria Europea, 2019). El uso de estos métodos de base científica no solo implica una mayor rapidez y menos costes, incrementando en gran medida la eficiencia y la simplificación de los procesos de los bancos (Carpi, 2023), sino que lleva estos procesos a un paso más, permitiendo alcanzar puntuaciones más precisas y a la eliminación tanto de falsos positivos como negativos, facilitando la determinación del plan de endeudamiento más adecuado para cada cliente (Federación Bancaria Europea, 2019).

En este contexto, la gran ventaja que plantea la IA surge cuando las entidades crediticias se topan con historiales crediticios incompletos. En efecto, estos sistemas aprovechan su capacidad para analizar *big data*, evaluando gran variedad de datos externos relacionados con el riesgo crediticio como pueden ser procesos judiciales, registros comerciales e industriales, opiniones públicas en medios autorizados, etc (Ridzuan et al., 2024). Este empleo por parte de la IA de datos alternativos es justamente lo que la hace revolucionaria pues, cuando nos encontramos con un problema no estructurado en el que se buscan respuestas no exactas, como puede ser la predicción sobre la probabilidad de que un individuo haga frente a sus obligaciones de pago, la realidad es que cuanta mayor cantidad de datos mejor será la predicción (Carpi, 2023).

Así, mientras que los modelos tradicionales de scoring únicamente hacen uso de historiales de crédito formales obtenidos a través del cliente y con su consentimiento para su procesamiento (Collado-Rodriguez, 2023), los sistemas de IA pueden analizar fuentes de datos desestructurados o poco estructurados, mejorando la precisión de las predicciones de incumplimiento de pago en préstamos (Palma, 2021).

1.2. Inclusión financiera

El hecho de que la IA abra el acceso al crédito para aquellos que por la vía tradicional, por no contar con este historial crediticio formal, no hubieran podido obtener financiamiento (Vlaicu, 2024), pone de relieve cómo además de mejorar la precisión de las predicciones, la IA también puede emplearse como herramienta de inclusión financiera (Carrascosa, 2024).

Concretamente, el empleo de la IA en la evaluación crediticia ha fomentado esta inclusión financiera por dos motivos principales. Por un lado, los modelos de *machine learning* permiten captar nueva información de los mismos datos (Alonso-Robisco & Carbó, 2023). Esto implica que son capaces de identificar mejor las relaciones no lineales entre variables para así mejorar la predicción del riesgo y permitir la identificación de solicitantes que, a pesar de parecer inicialmente de alto riesgo, en realidad presentan una probabilidad de impago menor a la estimada (Aldasoro et al., 2019).

Por otro lado, como consecuencia del acceso a datos que no se encontraban disponibles con anterioridad, la IA permite brindar oportunidades financieras a los aplicantes considerados como *credit invisible*, ayudando a aquellas poblaciones como los jóvenes o personas de renta baja que antes no podían acceder a servicios financieros (Griffith, 2003). Así, y como ha quedado demostrado en varios estudios, la combinación de emplear *big data* con modelos de *machine learning* ofrece ventajas significativas en la predicción de impagos y la apertura del crédito (Huang et al., 2020).

1.3. Eficiencia, productividad y reducción de costes

Por otro lado, la IA es una herramienta de gran ayuda a la hora de mejorar la eficiencia de los procesos bancarios, pues mitiga el error humano y aumenta la rapidez de las decisiones en la concesión de préstamos (Narang et al., 2024). Tanto es así que, según un estudio de McKinsey, las herramientas de IA pueden reducir el tiempo de aprobación de créditos en un 70% (Rebolledo, 2024). Esta eficiencia incluye entre otras cosas la automatización de procesos, elaboración de informes, gestión de reclamaciones, clasificación de documentos y extracción automatizada de datos (Federación Bancaria Europea, 2019).

En este sentido, los algoritmos son capaces de procesar grandes volúmenes de datos y realizar predicciones más precisas sobre la solvencia de los prestatarios, aumentando en gran medida la eficiencia asignativa en el mercado de crédito (Sadok et al., 2022) y pudiendo reducir los tiempos de procesamiento, mejorar las dinámicas de pago, reducir los gastos operativos y permitir la apertura de vías de recurso para los solicitantes (Narang et al., 2024).

Un ejemplo de esto es el sistema Contract Intelligence Network (COIN) de JPMorgan Chase, que emplea algoritmos de aprendizaje automático para analizar documentos legales como pudieran ser contratos de préstamos. De esta manera, se reduce el trabajo manual y el tiempo de procesamiento para así mejorar la eficiencia en la revisión de estos contratos y extracción de su información (Ridzuan et al., 2024). Otro ejemplo de esto es el uso de chatbots para guiar a los solicitantes en el proceso de solicitud de préstamos, agilizando el proceso mediante la respuesta a las dudas planteadas y verificación de la información proporcionada por los prestatarios (Krajka, 2024).

1.4. Monitoreo y gestión de cobro

Junto con la evaluación crediticia, el monitoreo y la gestión de cobro son elementos claves para asegurar la salud financiera de la cartera de préstamos. En este contexto, los sistemas de IA son capaces de diseñar estrategias de cobro personalizadas

que se basan en sistemas que buscan anticipar el comportamiento de los clientes para reducir la morosidad y mejorar la recuperación de los créditos (Rebolledo, 2024).

Así, la identificación temprana de los clientes vulnerables que están en riesgo, por ejemplo mediante señales como cambios en los patrones de pago o en el comportamiento financiero del prestatario, para ofrecerles, antes de que el problema se agrave, un plan de reestructuración personalizado o una renegociación de las condiciones crediticias adaptadas a su situación financiera, resulta de gran utilidad a la hora de reducir la morosidad (De Lucas, 2024). De esta manera, según estudios recientes, instituciones que han incorporado IA en sus estrategias de cobro han tenido una reducción en tasas de morosidad de hasta un 20% (Rebolledo, 2024).

1.5. Detección y prevención del fraude

La Inteligencia Artificial también es de gran ayuda con respecto a fines de seguridad. En efecto, esta inteligencia es capaz de detectar patrones inusuales y posibles actividades fraudulentas que puedan relacionarse con la delincuencia financiera (Narang et al., 2024). Los fraudes, tradicionalmente, han sido agrupados en dos categorías principales. Por un lado, el fraude externo, que tiene lugar fuera de la infraestructura del banco, como puede ser suplantación de identidad, ataques a los bancos o a sus clientes en relación con transferencias y, por otro, el fraude interno, acciones de los empleados como puede ser la filtración de información privada (Federación Bancaria Europea, 2019).

El objetivo del empleo de la IA reside en la capacidad de diferenciar acciones normales de actos fraudulentos. Algunos ya han pasado a denominarlo “el problema de decisión binaria” por el que, a partir de bases de datos que contienen una descripción de las características de diferentes eventos y sus estados, esto es, si los eventos son fraudulentos o inofensivos, la IA se entrena para tomar decisiones (Federación Bancaria Europea, 2019). Así, se busca que sea capaz de identificar la asociación que existe entre las características y los estados, es decir, qué elementos son los que encajan con una actividad fraudulenta, mejorando continuamente sus capacidades mediante el aprendizaje a partir de nuevos datos (Narang et al., 2024).

1.6. Experiencia del usuario

Por otro lado, y en relación con el desarrollo de productos, la IA es capaz de analizar el comportamiento de los clientes y conocerlos en mayor profundidad a raíz de los datos de las transacciones. Esto tiene como objetivo brindar al usuario una mejor experiencia, ofreciéndole productos que estén adaptados a sus necesidad y expectativas. Por ejemplo, los robo-asesores o *robo-advisors* son sistemas de IA que han sido diseñados para otorgar un asesoramiento automatizado a los clientes, ofreciendo aquellos productos que más se adaptan al prestatario en función de su perfil (Asociación Española de Banca, 2019).

De esta manera, mediante una mejora en la personalización de la oferta de crédito, la IA supone una mejora en la confianza y la lealtad de los clientes, a quienes se les responden preguntas, resuelven problemas (Narang et al., 2024) y proporcionan recomendaciones útiles que pueden incluir "sugerencias sobre cómo alcanzar objetivos financieros, ofrecer un ajuste en los límites de crédito en función de los datos de comportamiento u ofrecer recomendaciones de productos basados en necesidades" (De Lucas, 2024).

1.7. Cumplimiento normativo

Más allá de esto y en otro plano, el riesgo crediticio no es el único en el que la IA jugará un papel importante, pues este tipo de inteligencia también será de gran utilidad para una mejor gestión del riesgo interno en la concesión de créditos, pudiendo detectar desajustes y lagunas en el cumplimiento normativo (McKinsey & Company, 2024). En este sentido, la IA tiene la capacidad de automatizar el monitoreo de los riesgos, minimizando así el tiempo y los gastos asociados a los procedimientos manuales de cumplimiento y reduciendo el riesgo de multas y sanciones regulatorias (Narang et al., 2024).

Así por ejemplo, para asegurar el cumplimiento con las normas de conducta, la IA podría permitir a los bancos controlar mejor y aumentar la visibilidad de todos los datos poseídos, incluyendo soluciones para la clasificación de datos independientemente de la fuente: bases de datos, transacciones, correos electrónicos, archivos Excel, etc

(Federación Bancaria Europea, 2019), convirtiéndose la IA en un instrumento de gran valor a la hora de asegurar el cumplimiento de la normativa en el desarrollo de los servicios financieros y la concesión de préstamos (Carrascosa, 2024).

2. PRINCIPIOS ÉTICOS DE LA IA Y SU RELACIÓN CON LA CONCESIÓN DE PRÉSTAMOS BANCARIOS

Las aplicaciones descritas previamente muestran el enorme potencial que tiene la IA para optimizar la concesión de préstamos bancarios. Dicho esto, no debe olvidarse que, debido a que las decisiones de concesión de préstamos pueden afectar significativamente a la vida de las personas y generar consecuencias financieras de largo alcance, estas aplicaciones deben desarrollarse siempre dentro de un marco ético sólido.

Así, cuando hablamos de guías éticas para el desarrollo de la IA, son muchas las organizaciones que han extraído de los derechos fundamentales los principios éticos que deben configurar sus sistemas. Concretamente, la comisión de expertos de la UE, en su guía ética para el desarrollo de una IA confiable, destaca cuatro principios éticos que deben ser respetados por todos los actores involucrados. Para entender el esqueleto ético del uso de esta inteligencia en el ámbito bancario, es crucial analizar cada uno de estos principios y su aplicación a la concesión de préstamos bancarios (Comisión de expertos de la UE, 2019):

I. Respeto por la autonomía humana: Siempre que la IA interactue con humanos, será necesario que esta respete la autodeterminación de estos.

   En la concesión de préstamos, la IA debe complementar la evaluación sin sustituir la supervisión humana ni comprometer la autonomía del solicitante.

II. Prevención de daño: Los sistemas de IA no deberán causar, aumentar el daño o afectar de manera adversa a los humanos.

   En la concesión de préstamos, la IA debe evitar decisiones que perjudiquen injustamente a clientes vulnerables.

III. Justicia: Los sistemas de la IA deben ser justos en su desarrollo, despliegue y uso.

En la concesión de préstamos, la IA debe garantizar decisiones justas, evitando sesgos discriminatorios y permitiendo que los clientes impugnen resoluciones injustificadas.

IV. Explicabilidad: Los procesos de la IA deben gozar de transparencia, de manera que sus razonamientos puedan explicarse a aquellos que se ven afectados por sus decisiones.

En la concesión de préstamos, la IA debe ser explicable, permitiendo a los solicitantes comprender las razones detrás de la aprobación o rechazo y garantizando mecanismos de auditoría.

Dicho esto, en la práctica, garantizar que se cumplan estos principios no siempre es sencillo. Las decisiones tomadas por los sistemas de IA en la concesión de préstamos que vulneren estos ideales de autonomía, prevención de daño, justicia y explicabilidad, podrán derivar en resultados discriminatorios, opacos o que vulneren la privacidad de los usuarios. Así las cosas, en el siguiente capítulo veremos los diferentes desafíos éticos que surgen en la aplicación de la IA y como estos ponen en peligro el cumplimiento de los principios éticos.

CAPÍTULO II. ALGUNOS DESAFÍOS ÉTICOS DEL USO DE LA IA EN LA CONCESIÓN DE PRÉSTAMOS BANCARIOS

La predicción de la solvencia de un individuo mediante Inteligencia Artificial es un proceso complejo y no completamente estructurado. No existe una única respuesta matemática definitiva, sino un rango de posibles evaluaciones, lo que incrementa el riesgo de aparición de una serie de desafíos éticos que surgen, en gran parte, debido al uso masivo de datos (Carpi, 2023).

Si bien el sesgo humano en los programadores puede influir en el diseño de los sistemas de IA, este es un problema de ética humana, más que de ética de la IA. Por ello el foco de análisis de este capítulo no está en esas decisiones humanas, sino en los desafíos éticos que emergen cuando la IA toma decisiones sin intervención directa de un juicio humano, estos son: los sesgos algorítmicos, la explicabilidad y la protección de datos.

1. DISCRIMINACIÓN Y SESGOS ALGORÍTMICOS

La discriminación y los sesgos algorrítmicos estan ampliamente relacionados con los principios éticos de justicia y prevención de daños mencionados en el capítulo primero. El legislador, consciente de los riesgos que conllevan los sistemas de IA en la evaluación de solvencia y calificación crediticia, ha catalogado estos sistemas como de alto riesgo. En concreto, el Considerando 37 de la Ley de IA los reconoce como tales, en la medida en que determinan el acceso a "recursos financieros o servicios esenciales como la vivienda, la electricidad y los servicios de telecomunicaciones". En efecto, el legislador reconoce el potencial riesgo discriminatorio de su uso, dejando claro que "los sistemas de IA usados con este fin pueden discriminar a personas o grupos y perpetuar patrones históricos de discriminación, por ejemplo, por motivos de origen racial o étnico, género, discapacidad, edad u orientación sexual, o generar nuevas formas de efectos discriminatorios" (Parlamento Europeo y Consejo de la Unión Europea, 2024).

Así, aunque a menudo se sostiene que la IA es objetiva e imparcial debido a su dependencia de los datos y su aparente ausencia de juicios humanos, las evidencias indican lo contrario, apuntando recientes estudios a que el uso de esta tecnología ha aumentado "el riesgo de impactos redistributivos en clases sociales protegidas como religión, género o raza" (Fuster et al., 2022). En la práctica, múltiples casos han demostrado cómo los algoritmos pueden replicar e incluso amplificar sesgos sistémicos. En este sentido, algunas grandes empresas tecnológicas han enfrentado problemas con sistemas basados en IA. Amazon, por ejemplo, se encontró con dos problemáticas; la exclusión de códigos postales de mayoría negra y la penalización de currículums femeninos (Griffith, 2023).

Si bien estos casos no están directamente relacionados con la concesión de préstamos, como veremos, las mismas dinámicas pueden trasladarse al ámbito financiero. Los sistemas de IA diseñados para evaluar el riesgo crediticio pueden generar decisiones discriminatorias, pudiendo reforzar ciclos de exclusión financiera y debilitando la confianza en la aplicación de la IA en el sector bancario (Iturmendi, 2023). Estos sesgos pueden manifestarse en cualquier sistema de IA independientemente de que emplee tecnología *machine learning* o modelos más tradicionales. Así, aunque es cierto que en los sistemas sin *machine learning* la trazabilidad es mayor, lo cual facilita la identificación de la fase del proceso donde surge el sesgo, esto no elimina el problema (Carpi, 2023). Por ello, es fundamental analizar los factores que contribuyen a la aparición de estos sesgos para así poder mitigar sus efectos, debiendo distinguirse entre la discriminicación relacionada con los datos utilizados en la IA (sesgos derivados de la mala calidad de los datos y sesgos por asimetría de datos) y la relacionada con los propios procesos de la IA (sesgos de retroalimentación y sesgos de *proxy*).

1.1. El sesgo en los datos

Cuando hablamos de la IA en el sector financiero, el principal sesgo con el que nos podemos topar es el sesgo de los propios datos que se emplean, pues los sistemas de IA aprenden a tomar decisiones basándose en estos datos (Collado-Rodriguez, 2023). Los sesgos de datos pueden definirse como "una distorsión sistemática en los datos muestreados que compromete su representatividad" (Olteanu et al., 2019), pudiendo surgir el sesgo bien cuando los datos no reflejan la situación mundial o cuando, los datos reflejan el mundo, pero esta situación no es deseable (Stoyanovich et al., 2020).

La primera causa de Stoyanovich por tanto nos habla de la calidad de los datos. Lo cierto es que la capacidad de procesamiento de los sistemas de IA y la utilización de grandes cantidades de datos facilita la inclusión de datos de mala calidad ética (desactualizados, obtenidos extra legalmente, datos discriminatorios etc), no pudiendo los sistemas de IA corregir sus decisiones y prescindir de estos datos (Lousada, 2024). En ocasiones, la mala calidad puede derivar de una calificación previa de los datos hecha por humanos, que llevan a cabo una discriminación que no es detectada y

corregida por los algoritmos, y por tanto es perpetuada por los sistemas de IA (Aragüez, 2024).

Respecto a la segunda causa, en la que los datos reflejan la situación mundial pero esto no es deseable, el sesgo puede manifestarse en dos vertientes principales: los datos tradicionales de solvencia, como historial crediticio, ingresos y antecedentes financieros, y los datos alternativos, ya mencionados en el capítulo segundo, que incluyen información no convencional, como patrones de consumo e interacciones digitales (O'leary, 2013). En ambos casos, la abundancia de datos tiende a favorecer a un sector determinado de la población dejando en desventaja a quienes, por razones históricas o tecnológicas, generan menos información accesible para estos sistemas.

Por un lado, en relación con los datos tradicionales de solvencia, los sesgos pueden introducirse en los sistemas de IA a través de datos históricos que reflejan prejuicios humanos (López, 2024). En efecto, la imparcialidad humana ha existido en nuestra realidad mucho antes que la llegada de la IA y los datos históricos que se empleen en el entrenamiento de los sistemas de IA tenderán a reflejar estos sesgos humanos. Esto se debe a que los resultados de los algoritmos solo pueden ser tan buenos como los datos proporcionados como entrada; en otras palabras, *garbage in, garbage out* (Giralt, 2024).

No solo históricamente ciertos grupos han recibido menos oportunidades de acceso al crédito sino que en la actualidad persisten desigualdades estructurales que afectan a determinados grupos socioeconómicos, muchos de los cuales carecen de historiales financieros sólidos. Esta falta de datos genera una brecha que los sistemas de IA tienden a reforzar, perpetuando la discriminación en la concesión de préstamos (Iturmendi, 2023). Lo que, es más, el riesgo de que la IA continue la discriminación historica se ve agravado en la medida en que la naturaleza fundamental de la IA es la de ingerir grandes volúmenes de información para producir resultados que no pueden ser anticipados, especialmente cuando los datos de entrada están contaminados con sesgos (Griffith, 2023). El problema radica en que una predicción basada en una falta de datos puede provocar que se concedan préstamos a las personas incorrectas o que se nieguen a candidatos solventes simplemente porque su historial financiero es escaso o poco representativo. Esto lleva a resultados desfavorables y a que ese porcentaje de la

población nunca consiga construir los datos necesarios para que se les otorguen préstamos en el futuro, en resumidas cuentas, un círculo vicioso (Heaven, 2021).

Lo mismo ocurre con los datos alternativos, que a pesar de aumentar la información estadística, mejorar las predicciones e incluso dar lugar a una inclusión financiera, también es alto el riesgo de que ocasionen una discriminiación negativa (Collado-Rodriguez, 2023). Esto se debe a que este tipo de datos merecen de cierta crítica. En primer lugar, porque no puede establecerse una relación directa entre estos y la solvencia del individuo, incrementando considerablemente las posibilidades de error. Ejemplos de esto incluyen el uso de variables como la velocidad en que un sujeto recorre las condiciones generales de contratación como muestra de su nivel de responsabilidad o el tipo de aplicaciones que tenga instaladas en su teléfono (Hurley & Adebayo, 2016). En este sentido, puede resultar problemático que una decisión que ostenta tanta relevancia para la vida de los consumidores dependa de factores que poco tienen que ver con la situación económica individual del prestatario (Collado-Rodriguez, 2023).

En segundo lugar, porque la abundancia de los datos no tradicionales solo se produce respecto de determinados sectores de la población, en concreto aquellos que tiene acceso al mundo de la tecnología y desarrollan su actividad en esta. Según un estudio publicado por la OIT, la mayoría de datos empleados por la IA provienen de un grupo poblacional denominado *WEIRD*, esto es, poblaciones occidentales, educadas, industrializadas, ricas y democráticas (en inglés, *white, educated, industrialized, rich and democratic*) (Gmyrek et al., 2024). En contraposición, otros sectores de la población carecen de información virtual, afectando estos datos de manera negativa a la predicción de la IA respecto de sus candidaturas (Carpi, 2023). Así, tiene lugar el mismo problema; estos individuos son considerados como perfiles de alto riesgo, no porque realmente lo sean, sino porque los sistemas carecen de datos suficientes para evaluarlos correctamente, generándose un nuevo mecanismo de exclusión financiera.

Así las cosas, tanto para el caso de los datos tradicionales como alternativos, se da lo que se conoce como "la estadística contra las minorías" donde, aunque el algoritmo se haya programado de manera neutra, el sistema aprende que los grupos de personas sobre los que hay una sobreabundancia son la normalidad que debe

potenciarse, tratándose las características de aquellas personas que carecen de datos como la anormalidad a excluir (Todolí, 2024).

De esta manera, el problema no radica en la discriminiación en sí misma, en tanto que cualquier sistema de evaluación crediticia discrimina de alguna manera al clasificar a los solicitantes en categorías de riesgo. De hecho, un sesgo absolutamente nulo sería inviable e incluso perjudicial para el desarrollo de la IA (Griffith, 2023). El verdadero problema surge cuando esta discriminación se basa en datos incompletos o de mala calidad, afectando injustamente a determinados sectores de la población. Como consecuencia de esto, la brechas histórica y digital se convierten en una brecha crediticia en la que, por un lado, el sesgo histórico perpetúa la exclusión financiera basada en decisiones pasadas y, por otro, el sesgo tecnológico crea nuevas formas de discriminación al depender del acceso digital de los individuos. Las desigualdades existentes en la sociedad no solo se reflejan, sino que pueden amplificarse, dificultando aún más la inclusión financiera de ciertos grupos (López, 2024).

1.2. El sesgo de los procesos de la IA

Otro de los principales desafíos en materia de discriminación radica en que la IA no opera con reglas predefinidas, sino que aprende y ajusta sus decisiones en función de los datos que recibe. Así, por ejemplo, en el contexto de la concesión de préstamos, en lugar de programar un software con una fórmula fija para calcular el riesgo de impago, a la IA se le suministran una serie de datos que contienen información sobre antiguos clientes, incluyendo cuáles de ellos han incurrido en impagos (Carpi, 2023). De esta manera, la IA es capaz de aprender y mejorar por sí sola, identificando patrones y ajustando sus criterios de evaluación para mejorar su capacidad predictiva sin necesidad de un humano que programe los pasos que el algoritmo debe seguir para llegar a un resultado (Federación Bancaria Europea, 2019).

Por un lado, esto implica que los resultados sesgados ofrecidos por el sistema de IA (independientemente del origen de estos sesgos) se vuelve a integrar en el conjunto de datos que utiliza el propio algoritmo. Así, por ejemplo, si un sistema de IA excluye la concesión de un préstamo como consecuencia de un sesgo, aparte de perjudicar a la persona solicitante, confirmaría la aplicación de ese mismo sesgo a otras personas con

las mismas características, consolidando el sesgo mediante la multiplicación del mismo, lo conocido como sesgo de retroalimentación (Lousada, 2024).

Por otro lado, esta autonomía implica una serie de riesgos que justifican que los resultados de la IA siempre sean verificados, pues la capacidad de aprendizaje autónomo de la IA implica que esta puede establecer correlaciones inesperadas entre variables, generando decisiones que no siempre pueden ser anticipadas ni explicadas fácilmente (Carpi, 2023). En relación con la discriminación, este fenómeno conduce a un problema clave: el uso de *proxies* discriminatorios. En efecto, los sistemas de decisión basados en IA operan con un alto nivel de complejidad y recogen una pluralidad de variables, siendo capaces de identificar factores estructurales dentro de la sociedad y usarlos como sustitutos de características protegidas (raza, género, edad etc) (Lousada, 2024).

En el contexto de la concesión de préstamos bancarios, los datos alternativos son vulnerables a ser empleados por los algoritmos de IA como *proxies* para la discriminación, incluso si han sido programados explícitamente para no considerar factores prohibidos. Es decir, aunque la IA no utilice directamente datos como la raza o el género para determinar el riesgo crediticio, sí podrá usar otros datos que están altamente relacionados con estas características y empezar a considerarlos relevantes para la determinación de la solvencia de los candidatos (por ejemplo, código postal, patrones de gasto, estudios) (Kou et al., 2019). Este fenómeno dificulta la detección del sesgo, ya que no se trata de una discriminación explícita, sino de un patrón que emerge del aprendizaje automático. En este sentido, los *proxies* o sustitutos son difíciles de identificar y remediar, siendo un desafío importante para los programadores decidir qué factores excluir, pues estos pueden hacer juicios intuitivos e imprecisos que no tengan en cuenta el potencial de determinadas variables como sustitutas de características prohibidas (Griffith, 2023).

Por todo esto, el uso de proxies en los sistemas de IA plantea un desafío ético fundamental pues la IA, al operar con un aprendizaje automático, puede desarrollar patrones discriminatorios sin que estos sean evidentes a simple vista, dificultando la detección y corrección del sesgo, e impidiendo que los afectados puedan comprender las decisiones tomadas en su contra. En efecto, corregir un sesgo requerirá que se conozca

cómo el proceso está sesgado (Stoyanovich et al., 2020), lo que nos lleva al siguiente punto: la explicabilidad.

2. EXPLICABILIDAD Y RENDICIÓN DE CUENTAS

Como ya se analizaba en el apartado de los principios éticos, la explicabilidad es uno de los pilares fundamentales necesarios para construir una IA confiable pues, sin la capacidad de comprender y justificar las decisiones de un sistema automatizado, la transparencia y la rendición de cuentas pueden verse comprometidas. Respecto a cómo podemos definir y entender la explicabilidad, existen varias propuestas. Así, por ejemplo, el Reglamento General de Protección de Datos de la Unión Europea lo define como la "divulgación significativa de la lógica subyacente". Por su parte, el Banco de Inglaterra, lo define como un sistema en el que el "interesado pueda comprender los principales factores que impulsan una decisión basada en un modelo" (Joseph, 2020). Quizás, la definición más completa es aquella brindada por el Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial, que define la explicabilidad como "la capacidad de explicar tanto los procesos técnicos de un sistema de IA como las decisiones humanas relacionadas (por ejemplo, las áreas de aplicación de un sistema)". De esta forma, la explicabilidad técnica requiere que "las decisiones tomadas por un sistema de IA puedan ser comprendidas y rastreadas por los seres humanos" (Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial, 2019).

Partiendo de estas definiciones, en la práctica, la explicabilidad consiste en hacer comprensible tanto los procedimientos de creación de los algoritmos y su relación con los resultados obtenidos como el funcionamiento del propio algoritmo, permitiendo que se de una verdadera trazabilidad del proceso (Megías, 2022). Garantizar la explicabilidad de estos procesos permite que las decisiones algorítmicas sean revisadas, comprendidas y, en caso necesario, impugnadas. De esta forma, aquellos que pudieran verse negativamente impactados por las decisiones tomadas por la IA podrán conocer los motivos detrás de la decisión automatizada y emprender las acciones de carácter jurídico oportunas (Ruiz, 2022).

Por el contrario, cuando los sistemas de IA carecen de explicabilidad, no es posible justificar de manera clara cómo y por qué toman una decisión, no solo

dificultando que los afectados puedan cuestionar decisiones que les perjudican, sino también impidiendo que se identifiquen y corrijan posibles sesgos en los modelos (Lousada, 2024). Para mitigar esta falta de explicabilidad, es necesario comprender cómo y por qué surge, siendo el objeto de este capítulo abordar el origen de este problema ético así como su impacto en los modelos de aprendizaje automático.

2.1. La opacidad en los algoritmos

En primer lugar, uno de los principales problemas que afecta a la explicabilidad de la IA radica en el procesamiento de datos. En muchas ocasiones, los algoritmos de IA dependen de información que pasa por múltiples actores e intermediarios, lo que dificulta rastrear quién ha intervenido en cada fase del proceso (Barrat, 2023). En este contexto, pueden darse parcialidades en la recabación de los datos o un traslado de los sesgos de los programadores al algoritmo, pudiendo quedar estos sesgos cubiertos trás una cortina de opacidad (Felzmann et al., 2020). Esta opacidad es la antítesis de la explicabilidad y deriva de la existencia de las denominadas "cajas negras", esto es, agujeros negros digitales en los que tienen lugar procesos de decisión que carecen de capacidad explicativa y resultan ininteligibles para quien las recibe (Rivas, 2020). Con esto en mente, la opacidad genera dos imaginarios populares, por un lado aquellos que tienden a confiar en la perfección del algoritmo y aceptar de manera acrítica sus decisiones (Goñi, 2019) y, por otro, aquellos que pierden la confianza respecto de las novedades que pueda aportar la IA, provocando una desincentivación así como una gran cantidad de litigios y demandas en el ámbito legal (Ruiz, 2022).

En este sentido, cabe mencionar que la explicabilidad no busca la revelación de los códigos empleados en el software de los sistemas de IA sino que, en la medida de lo posible y de forma razonable, puedan explicarse las decisiones tomadas por el algoritmo y por tanto pueda tener lugar una rendición de cuentas (Cortina, 2019). Dicho de otro modo, la finalidad de esta explicabilidad no es otra que permitir la identificación de qué o quién es responsable de sus resultados. El desafío por tanto radica en que, a diferencia de los sistemas tradicionales, en los que las decisiones pueden atribuirse directamente a un individuo, en la IA no encontramos un sujeto humano detrás de cada decisión individual, sino una máquina que carece de personalidad y a la que es difícil exigir responsabilidad (Colmenarejo, 2018). Así, la opacidad causa que "los sesgos y los

responsables de haber adoptado las decisiones correspondientes queden ocultados detrás de una fórmula matemática" (Rodriguez, 2024).

Así, el riesgo más grave de esta opacidad se manifiesta cuando los programadores diseñan algoritmos de creciente complejidad, hasta el punto en que la comprensión del algoritmo deviene imposible, incluso para el propio creador (Ruiz, 2022). Mientras que los softwares tradicionales previos a la IA funcionan bajo reglas claras, emitiendo un resultado que se limita a lo programado y por tanto es fácilmente predecible, la IA de aprendizaje automático puede hacer combinaciones de datos muy variados para desarrollar nuevas correlaciones y patrones de decisión que no han sido ideados o previstos por sus creadores sino aprendidos por el sistema posteriormente (Carpi, 2023).

Concretamente, en el ámbito de la concesión de créditos, esto adquiere una dimensión especialmente crítica. En efecto, cuando el algoritmo opaco toma una decisión, ésta deviene una especie de veredicto inapelable, ya que ni los afectados ni las propias entidades financieras pueden comprender con precisión qué factores han influido en el rechazo o aprobación del crédito (datos parciales, incorrectos, malinterpretados…), siendo fundamental que los responsables de la gestión bancaria pueden motivar en que se basan sus decisiones y proporcionar explicaciones claras a los clientes (Funcas, 2024). Esto mismo señala el Reglamento General de Protección de Datos de la Unión Europea, el cual exige que las entidades financieras proporcionen explicaciones que justifiquen la denegación de los créditos o préstamos (RGPD, 2016). De no hacerlo y permitir que la falta de explicabilidad se convierta en una nueva normalidad, nos enfrentamos al riesgo de construir una sociedad excesivamente tecnificada, en la que incluso las elites tecnocráticas que dominan estos algoritmos desconozcan los procesos que estos llevan a cabo y no sea posible encontrar un sujeto responsable de los resultados de las decisiones automatizadas (Coekelbergh, 2020).

2.2. Explicabilidad y *machine learning*

En este contexto, la lucha contra la opacidad deviene difícil en aquellos modelos de IA basados en aprendizaje automático o *machine learning*. Como ya ha sido analizado *supra,* con esta tecnología, el algoritmo aprende, inicialmente, a partir de los

datos suministrados y, posteriormente, en el análisis, procesamiento y combinación de estos datos para finalmente elaborar una decisión que resuelva el problema planteado (Carpi, 2023). En otras palabras, son sistemas capaces de aprender de los datos suministrados y ajustar su comportamiento de forma autónoma, de manera que aunque sea posible saber cómo funciona el sistema en general no sea posible explicar una decisión particular. Así, a medida que estos algoritmos se van modificando y aprenden por su cuenta, su lógica interna se vuelve cada vez más opaca, generando patrones de decisión que ni sus propios creadores comprenden (Coekelberg, 2020).

Por todo esto, aunque las técnicas de aprendizaje automático de IA son las más prometedoras en el ámbito de la concesión de créditos, pues como vimos en el capítulo primero su uso podría dar lugar a unas predicciones más precisas, también son las que tienen un nivel más bajo de explicabilidad (Federación Bancaria Europea, 2019). Tanto es así, que la Autoridad Bancaria Europea autoriza en menor cantidad el uso de algoritmos de aprendizaje automático, reconociendo su elevada complejidad en términos de justificación y supervisión humana en el proceso de toma de decisiones (Carrascosa, 2024). Dicho esto, aunque la mayoría de sistemas de perfilado crediticio no dependen exclusivamente de tecnología *machine learning*, la tendencia del futuro será avanzar hacia estos sistemas de aprendizaje automático. Por ello, el *quid* de la cuestión, que se abordará en el capítulo tercero, consistirá en cómo mantener una buena precisión predictiva sin poner en peligro la interpretabilidad de los sistemas de IA, garantizando así que la automatización no sacrifique la explicabilidad ni la rendición de cuentas.

3. PROTECCIÓN DE DATOS

Otro de los grandes problemas generados por la aplicación de la IA en la gestión del riesgo de crédito surge en el terreno de la protección de datos que goza de una estrecha relación con los principios éticos de respeto por la autonomía humana y prevención de daño. Aunque este tipo de riesgo ya existía y suponía un desafío legal en los sistemas de calificación crediticia tradicionales, con los sistemas automatizados se ha visto incrementado significativamente (Carpi, 2023).

En este sentido, el artículo 22 del Reglamento General de Protección de Datos (RGPD) establece que, en aquellos casos en los que una actividad pueda generar efectos legales sobre los individuos o afectarles de forma significativa, se prohíbe que estas decisiones sean tomadas únicamente por procesos automatizados, incluyendo la creación de perfiles crediticios (RGPD, 2019). Las excepciones a esto son que haya consentimiento del individuo; que el proceso automático que genera la calificación crediticia sea necesario para formalizar un contrato entre el individuo y el gestor de los datos, o que esté autorizado por una ley nacional y se pongan salvaguardias (Carrascosa, 2024).

En este contexto la justicia es clara, restringiendo los procesos automáticos y prevaleciendo los derechos a la intimidad de los individuos. Así, la sentencia del 7 de diciembre de 2023 del Tribunal de Justicia de la Unión Europea (TJUE) interpreta el artículo recalcando que los bancos no podrán contratar proveedores externos de calificación crediticia que usen procesos automáticos para generar las calificaciones. Así, sólo se permitirá el uso de estos procesos automáticos de forma interna, debiendo el particular recibir una información previa si su solvencia se va a basar en procesos automáticos (TJUE, 2023).

Partiendo de este marco legal establecido por el RGPD y reforzado por la jurisprudencia de la Unión Europea, que impone límites claros al uso del aprendizaje automático en la evaluación del riesgo crediticio, es fundamental analizar qué riesgos éticos surgen a raíz del tratamiento de datos personales, siendo estos riesgos de naturaleza variopinta y de desarrollo en distintos planos.

3.1. Principios de minimización de datos y limitación a la finalidad

En primer lugar, cabe destacar que más allá del marco legal analizado *supra*, el Reglamento General de Protección de Datos señala en su artículo 5 un conjunto de principios que los responsables del tratamiento deberán observar al tratar datos personales, destacando los principios de minimización de datos y limitación de la finalidad (RGPD, 2019). El principio de minimización de datos establece que solo deben recopilarse y tratarse aquellos datos estrictamente necesarios, es decir, aquellos que sean “adecuados, pertinentes y limitados” para cumplir con un propósito específico,

evitando la recolección excesiva de información. El principio de limitación de la finalidad, por su parte, exige que los datos personales sólo sean utilizados para los fines "determinados, explícitos y legítimos" para los cuales fueron recogidos, prohibiendo su uso posterior para objetivos distintos sin el consentimiento del titular (Agencia Española de Protección de Datos, 2023).

Estos dos principios suponen un reto para el desarrollo y empleo de la IA. Por un lado, la minimización de los datos puede entrar en conflicto con la IA en la medida en que esta requiere de una gran cantidad de volumen de datos para generar predicciones más precisas (Federación Bancaria Europea, 2019). En efecto, las tecnologías de IA y sobre todo los algoritmos de aprendizaje automático aprenden "con el ejemplo", por lo que cuantos más datos sean proporcionados al sistema, mejor podrán aprender y aplicar sus conclusiones a datos aún no vistos (Bishop, 2006).

Por otro lado, la limitación a una finalidad también puede ser difícil de cumplir en tanto que entra en conflicto directo con la naturaleza de la IA, que busca procesar los datos más allá de su propósito inicial para así mejorar su capacidad de predicción y análisis (Barrat, 2023). En concreto, este reciclaje ha generado preocupaciones en relación con la posibilidad de que datos sensibles sean almacenados y reutilizados sin supervisión, corriendo las entidades el riesgo de perder el control sobre el almacenamiento y distribución de los datos (Cloudfare, 2025).

De seguirse estos dos principios, la IA perdería gran parte de su funcionalidad, por lo que, hasta que el RGPD contenga disposiciones específicas sobre la IA más adaptadas a su naturaleza, estos desafíos continuarán existiendo (Barrat, 2023).

3.2. El uso de datos personales alternativos

Por otro lado, el uso de datos alternativos por parte de la IA también supone un reto en el ámbito de la protección de datos, siendo cada vez más las entidades financieras que recurren a estos datos para la evaluación de solvencia y la concesión de créditos. Aunque de manera aislada este tipo de datos no revelen nada sobre la solvencia de su titular, al ser tratados estadísticamente pueden revelar patrones sobre los comportamientos crediticios habituales de un determinado grupo, de manera que sean

útiles para la predicción del futuro comportamiento financiero del individuo (Carpi, 2023). El uso de este tipo de datos plantea diversos riesgos éticos, de entre los cuales destacamos tres en concreto.

En primer lugar, el artículo 16 RGPD recoge el derecho del interesado a la rectificación de los datos personales inexactos. Esta disposición tiene implicaciones de gran relevancia en el contexto de la IA, pues al tomar decisiones basadas en cantidades de datos de diversas fuentes, que no han sido corroboradas, existe el riesgo de que se propaguen las inexactitudes, dando lugar a decisiones incorrectas (Barrat, 2023). Así, es cuestionable que los sistemas de IA que analizan *big data* toleren cierta cantidad de datos desordenados o inexactos como consecuencia de que los volúmenes de datos procesados suelen ser muy grandes, pues estas inexactitudes podrían llevar a predicciones incorrectas y afectar de forma negativa a los interesados (Mitrou, 2018).

En segundo lugar, el Supervisor Europeo de Protección de Datos (*EDPS*, en inglés) establece que "aunque los datos sean públicos en la red, eso no significa que su tratamiento sea legal", pues los prestatarios no tienen porqué haber manifestado un consentimiento a que esos datos alternativos sean usados para el análisis de su solvencia. Esto plantea un problema ético que la guía de la EBA trata de solucionar estableciendo que, para la perfilación del cliente, deberán usarse datos que guardan una clara relación con la capacidad de solvencia, aunque, en la práctica, esto no será fácil de cumplir (Carrascosa, 2024).

Por otro lado, el uso de estos datos alternativos, obtenidos sin el consentimiento del individuo y que en muchas ocasiones guardan poca relación con su capacidad económica, de no controlarse su tratamiento, podrá dar lugar a resultados discriminatorios. En este sentido, su uso está fuertemente restringido, previéndose en la UE y en España un marco normativo estricto que no permite el empleo de buena parte de estos datos para las evaluaciones de solvencia (Carpi, 2023), quedando prohibido el procesamiento de datos de categorías especiales relativos al "origen étnico o racial, las opiniones políticas, las convicciones religiosas o filosóficas, o la afiliación sindical, la vida o la orientación sexuales" aunque con determinadas excepciones (RGPD, 2019).

Sin embargo, como ya se comentaba en el apartado del sesgo algorítmico, existe un problema adicional, esto es, el uso de *proxies*. Muchas entidades recopilan datos de comportamiento en línea, interacciones en redes sociales y hábitos de consumo sin que el usuario sea plenamente consciente de ello. Estas variables, aunque no forman parte de las estrictamente protegidas, sirven como sustitutos de estas características, siendo la recabación no controlada de estos datos un reto ético, pues genera riesgos de discriminación sin que se infrinja técnicamente la normativa (Bahoo et al., 2023).

3.3. Seguridad de la IA

Por último, en la medida en que la IA aún se encuentra en sus primeras etapas de desarrollo, los fallos en la seguridad resultan más difíciles de corregir, siendo común que tengan lugar muchos problemas técnicos y riesgos. Al ser una tecnología en constante evolución, los sistemas de IA todavía presentan vulnerabilidades difíciles de prever, lo que los convierte en objetivo de ciberataques (Ayerbe, 2020). Uno de los mayores riesgos en este sentido consiste en que los delincuentes puedan hacer uso de tecnologías relacionadas para dirigir ataques contra el sistema y obtener acceso a los datos personales de los clientes para cometer delitos (Ridzuan et al., 2024).

En esta misma línea, existe un riesgo relacionado con la capacidad de la IA para eludir los controles de privacidad de datos. Investigaciones recientes han revelado que es posible manipular modelos de IA para que revelen información personal, como se demostró con un fallo en el software de IA de Nvidia, donde investigadores lograron engañar al sistema para obtener datos protegidos, lo cual plantea serias dudas sobre la capacidad de las organizaciones para garantizar la seguridad de la información cuando dependen de herramientas de IA (Cloudfare, 2025)

Por último, cabe mencionar que los sistemas de IA son especialmente vulnerables a los denominados ataques adversariales, donde el objetivo es alterar el comportamiento del modelo. Así, técnicas como los ataques de inversión de modelo, que permiten reconstruir información sensible, o los ataques de inferencia de pertenencia, que revelan si una persona específica formaba parte del conjunto de entrenamiento, permiten deducir información personal sensible y ponen en peligro la protección de datos (Xi, 2021).

CAPÍTULO III. POTENCIALES MEDIDAS DE ATEMPERACIÓN DE LOS PROBLEMAS ÉTICOS

1. MITIGACIÓN DEL SESGO ALGORÍTMICO

Para abordar la problemática relativa a los sesgos que se producen en el uso de la IA para la concesión de préstamos, se han desarrollado diversas estrategias de mitigación. Si bien es cierto que eliminar por completo el sesgo en la IA es prácticamente imposible, pues existen sesgos inherentes a la condición humana, hay una serie de buenas prácticas que permiten mitigar los efectos que estos sesgos producen. Concretamente, para mitigar los sesgos introducidos en el capítulo tercero, esto es, sesgo por mala calidad y por asimetría de datos, y sesgos de *proxy,* las estrategias pueden agruparse en dos niveles, por un lado, el tratamiento de los datos y resultados generados por la IA y, por otro lado, el plano humano en la supervisión y control del sistema.

1.1. Estrategias de tratamiento de datos

1.1.1. Tratamiento previo al entrenamiento de la IA

Desde el primer enfoque, una de las estrategias más relevantes para la mitigación del sesgo algorítmico, consiste en el preprocesamiento de los datos previos al entrenamiento de la IA. Esta técnica se basa en garantizar que los datos utilizados para entrenar los modelos de IA sean representativos de la población, evitando la introducción de sesgos desde la fase inicial. Para lograrlo, las entidades de crédito deben priorizar el uso de fuentes de datos confiables y de alta calidad a la hora de entrenar sus modelos, evitando en la medida de lo posible conjuntos de datos que puedan dar lugar a sesgos (Banco de España, 2024).

En este sentido, existen varias técnicas que pueden aplicarse para mejorar la representatividad de los datos y mitigar los sesgos. Las principales son el rebalanceo de datos (*re-weighting*) y el remuestreo de datos (*resampling*). Por un lado, el *re-weighting* consiste en transformar los datos mediante la modificación de los pesos que a estos se

les asignan. De esta manera, se asignan los pesos más bajos a aquellos grupos poblacionales que tienen una mayor probabilidad de obtener resultados favorables, otorgando a aquellos con menos probabilidades un peso más alto, evitando así la sobrerrepresentación de ciertos sectores de la población (Seiffert et al., 2008).

Estudios llevados a cabo demuestran cómo esta técnica funciona para mitigar el sesgo de los modelos de puntuación crediticia, garantizando una mayor equidad en la asignación de créditos. Por otro lado, el resampling consiste en la modificación del tamaño del conjunto de datos que se emplean (sobremuestreo o submuestreo), afectando a la distribución sin tener que transformar los datos ni modificar los pesos asignados. En la mitigación de sesgos, la técnica más frecuente será el sobremuestreo, siendo las dos principales estrategias la *Synthetic Minority Over Sampling Technique* (SMOTE) y las Redes Generativas Adversarias (GANs) (González-Sendino et al., 2024).

En concreto, las GANs llevan a cabo una generación de datos sintéticos para solucionar la falta de datos representativos de ciertos grupos poblacionales. Es decir, pueden crear datos artificiales basados en patrones reales, reduciendo la subrepresentación de ciertos grupos en los datos de entrenamiento para ayudar a mejorar la equidad en los modelos sin comprometer la privacidad de los individuos (Xu et al., 2019).

1.1.2. Tratamiento durante el entrenamiento de la IA

Además del preprocesamiento de datos, existen estrategias que pueden implementarse durante la fase de entrenamiento del modelo que buscan limitar la discriminicación que la IA puede aprender de los propios datos. Por un lado, la regularización es una técnica que busca reducir las disparidades que puedan ocasionarse entre las predicciones de los distintos grupos. Esto se logra mediante la penalización de las altas correlaciones entre atributos sensibles (como género o raza) y los resultados del modelo. Dicho esto, una de las grandes desventajas de este tipo de método reside en que puede empeorar la explicabilidad de los resultados (Stevens et al., 2020). De manera similar, estrategias como el *Fairness-Aware Learning* consisten en mecanismos de ajuste dinámico, donde los algoritmos modifician sus pesos en tiempo real en caso de detectar discriminación en sus predicciones (Zhang et al., 2018)

Por su parte, la estrategia de eliminación adversarial de sesgo (*adversarial de-biasing*) consiste en el entrenamiento simultáneo de dos redes neuronales. De esta forma, una de las redes neuronales se encarga de predecir la solvencia crediticia mientras que la otra se encarga de eliminar los sesgos en los datos de entrenamiento que podrían afectar a la predicción (Zhang et al., 2018). Realmente lo que ocurre es que se enfrentan dos modelos de *machine learning*, el primero predice la solvencia crediticia y el segundo intenta identificar raza, género u otros atributos protegidos del solicitante que evalúa el modelo que predice. Estos dos modelos compiten entre ellos, de ahí la palabra "adversarial", hasta que el segundo modelo no es capaz distinguir la raza o el género en las salidas del primer modelo, dando lugar a un modelo final que es preciso y a la vez justo (Zest AI, 2019).

1.1.3. Tratamiento posterior al entrenamiento de la IA

Por último, la discriminación aprendida por la IA podrá corregirse mediante la modificación de la salida del modelo, existiendo dos algoritmos principales. Por un lado, *Equalized Odds*, un algoritmo que ajusta las probabilidades de los resultados finales con un objetivo de igualdad de oportunidades y, por otro, la Clasificación con Opción de Rechazo (*Reject Option Classification*), en la que en los casos de mayor incertidumbre se favorece a los grupos subrepresentados en la asignación de resultados favorables, mientras que a los grupos privilegiados se les asignan resultados menos favorables (González-Sendino et al., 2024).

1.2. Estrategias de supervisión y control humano

En el plano humano, las estrategias para la prevención de sesgos en la toma de decisiones consisten en que los modelos de IA sean monitoreados, evaluados y ajustados de forma rutinaria (Ridzuan et al., 2024), no pudiendo olvidar la necesidad de una colaboración entre humanos y máquinas para alcanzar resultados justos y eficientes (Rizinski et al., 2022). Es decir, adoptar un enfoque denominado "humano en el circuito" o *human-in the loop* hace que los humanos tengan control sobre los sistemas y puedan determinar si las decisiones tomadas por estos son seguras. En este sentido, se pueden identificar dos maneras de interacción humana en el empleo de la IA.

En primer lugar, el sistema ya mencionado “humano en el circuito”, donde el sistema de IA únicamente hará sugerencias que deberán ser evaluadas por humanos que finalmente tomarán la decisión. La segunda, “humano supervisando el circuito” o *human over the loop*, donde el humano supervisa y monitorea el sistema de IA en caso de que sea necesaria su intervención, por ejemplo, porque la IA comienza a mostrar patrones de comportamiento inesperados o inadecuados. Para facilitar la auditoría de estos modelos, existen pruebas de equidad y herramientas como AI Fairness 360 que permiten evaluar el impacto de la IA en distintos grupos poblacionales. Añadiendo este factor humano, es posible mitigar resultados discriminatorios a los que pueda llegar la IA (Ridzuan et al., 2024).

Sin embargo, cabe mencionar que este mecanismo sólo será realmente efectivo si los humanos que supervisan los sistemas pueden comprender e interpretar las decisiones tomadas por la IA. En otras palabras, si el algoritmo empleado tiene una baja explicabilidad, los humanos que interaccionan con este no serán capaces de detectar estos sesgos algorítmicos, siendo necesario que este factor humano vaya acompañado también de una interpretabilidad que permita comprender las decisiones algorítmicas (Doshi-Velez, 2017). Respecto a las estrategias que pueden llevarse a cabo para obtener esta explicabilidad, nos adentraremos en esta problemática en el siguiente apartado.

2. MEJORA DE LA EXPLICABILIDAD Y RENDICIÓN DE CUENTAS

Como ya se comentaba en el capítulo primero, los modelos de IA de aprendizaje automático, como por ejemplo las redes neuronales, ostentan un poder predictivo de mayor envergadura. Sin embargo, este tipo de IA funciona a menudo como cajas negras, generando resultados sin una clara explicación detrás de sus decisiones, lo que pone en peligro el principio ético de la explicabilidad. La dificultad de interpretar estos modelos plantea un dilema: ¿es preferible utilizar modelos de IA con alta capacidad predictiva, aunque sean opacos, o sacrificar precisión en favor de la transparencia? Que el principio de explicabilidad quede comprometido no parece razón suficiente para impedir por completo el uso de este tipo de IA, pues su uso aporta una ventaja competitiva significativa para las entidades bancarias. Así, será necesario acudir a estrategias para la mitigación de esta falta de explicabilidad (Alonso-Robisco et al., 2023).

2.1. La Inteligencia Artificial Explicable

Una posible solución a este problema es el desarrollo de un sistema híbrido de dos niveles, en el que modelos de IA altamente predictivos pero con baja explicabilidad se complementan con técnicas de análisis explicativo que proporcionen el nivel de transparencia requerido (Federación Bancaria Europea, 2019). Así, la principal solución para la mitigación de este riesgo es la utilización de IA explicable (en inglés, XAI o *Explainable AI*), que permite que el uso de algoritmos *machine learning* en el análisis de riesgo de lugar a resultados claros y justificados (Hoffman et al., 2018). Este tipo de IA está ligada a una serie de conceptos que contribuyen a una mayor explicabilidad. En primer lugar, la intervención y supervisión humana, en segundo lugar, la transparencia entendida como la comunicación adecuada a los interesados de las decisiones tomadas por la IA y, por último, la responsabilidad, debiendo los sistemas de IA ofrecer mecanismos que garanticen que exista una rendición de cuentas, auditoría y evaluación de los algoritmos (Nallakaruppan et al., 2024).

Así, esta IA explicable está integrada por estrategias de interpretación de modelos que buscan solucionar las dificultades planteadas por la opacidad, siendo las tres más conocidas SHAP (Lundberg, 2017), LIME (Ribeiro, 2016) y FI (*Permutation Feature Importance*) (Breiman, 2001). En concreto, LIME es una de las pruebas de explicabilidad más utilizadas, principalmente por su carácter neutral en relación con el modelo. Este test realiza una explicación local de las decisiones de la IA (Wachter et al., 2017) a base de utilizar datos tomados en una ubicación cercana a la de los datos relevantes, para así construir un modelo interpretable local que emplea un comportamiento lineal y por tanto es más fácil de comprender por las personas (Alonso-Robisco & Carbó, 2023). Cómo LIME no depende del modelo y puede generar explicaciones para cualquier sistema de aprendizaje automático o *machine learning,* es capaz de lograr una mayor interpretabilidad que otros sistemas (Nallakaruppan et al., 2024). En otras palabras, se trata de combinar los modelos tradicionales, como la regresión lineal, con modelos de IA para así poder mejorar la explicabilidad de las decisiones tomadas (Bahoo et al., 2023)

2.2. Restricción de la monotoneidad

Otra alternativa para mejorar la explicabilidad sin sacrificar la precisión de los modelos de IA es el desarrollo de modelos que combinen la capacidad predictiva del *machine learning* con restricciones que garanticen una mayor transparencia. Concretamente, imponer la monotoneidad entre las características y las salidas del modelo mejora la explicabilidad del sistema, en la medida en que permite entender el efecto de una única característica sobre la salida del modelo de manera independiente a otras características (Nguyen, 2019). En el ámbito de la concesión de préstamos bancarios, esto implica que si una variable relevante para la concesión de crédito aumenta, la probabilidad de aprobación del préstamo debería aumentar o disminuir de manera consistente, evitando decisiones contraintuitivas y facilitando las explicaciones a reguladores y clientes (Banco de España, 2022).

En relación con esto, cabe destacar un informe del Banco de España sobre el proyecto NeuroDecision Technology (NDT), en el que se buscaba mejorar la capacidad predictiva de los modelos tradicionales como la regresión lineal y a su vez reducir la opacidad de los modelos de *machine learning*, mitigando el problema de la "caja negra" y mejorando la trazabilidad de las decisiones. En este estudio se empleaba un modelo NDT, basado en *machine learning* con restricción de la monotoneidad. Al ser más transparente que modelos tradicionales de *machine learning* (redes neuronales, Random Forest etc), NDT permitió un mayor control y auditoría de las decisiones sin sacrificar la predicción (aunque no logró avance en este ámbito). Así, la implementación de modelos con restricciones de monotoneidad mejorarían la trazabilidad, permitiendo una mejor auditoría y cumplimiento normativo (Banco de España, 2022).

Sea cual fuere la estrategia que se decida implementar, la Federación Bancaria Europea defiende un enfoque basado en el riesgo, proporcionado distintos niveles de explicabilidad en función del riesgo de la situación concreta, yendo desde un alto grado de explicabilidad hasta una funcionalidad probada. Así, el nivel de detalle facilitado se basaría en el impacto de los resultados del sistema para las personas (Federación Bancaria Europea, 2019).

Por último, también será fundamental el seguimiento y la actualización de los modelos, siendo necesario que la entidades estén constantemente mejorando los modelos para evitar que estos desarrollen sesgos ocultos o pierdan precisión con el tiempo, Para ello, es fundamental contar con con equipos de expertos en ciencia de datos capaces de manejar la complejidad de estos sistemas y garantizar su correcto funcionamiento (Funcas, 2024).

3. PROTECCIÓN DE DATOS Y PRIVACIDAD

Como ya quedó mencionado en el capítulo anterior, uno de los grandes retos éticos del aprendizaje automático consiste en el empleo de una gran cantidad de datos de entrenamiento, cuyo uso puede infringir el Reglamento General de Protección de Datos. Para mitigar estos riesgos, existen diversas estrategias diseñadas para garantizar la privacidad y seguridad de la información.

3.1. Anonimización y seudonimización de los datos

En primer lugar, una de las principales estrategias para proteger la privacidad de los datos es el uso de técnicas de anonimización y la seudonimización. En efecto, conforme al artículo 5.1.b RGPD y el considerando 156, los datos recabados por las entidades de crédito podrán ser tratados siempre que se anonimicen o seudonimicen (RGPD, 2016). Mientras que la seudonimización es una técnica reversible que consiste en disfrazar las referencias personales mediante la eliminación de identificadores directos y el uso de seudónimos en su lugar, la anonimización consiste en la eliminación de toda aquella información que permita una identificación, de manera que dicha información no pueda ser rastreada ni reconstruida (Agencia Española de Protección de Datos, 2021).

Ahora bien, en el contexto del RGPD, si se busca evitar la aplicación contundente de esta normativa, tendremos que recurrir a la anonimización, pues la seudonimización continúa estando sujeta en gran medida a esta regulación; debiéndose cumplir determinadas obligaciones entre las cuales se incluye la eliminación de datos una vez no ya no haya obligación de conservación ni motivo para mantenerlos. Esto se debe a que la finalidad del tratamiento de los datos deberá ser equivalente a la finalidad

de la obligación legal a la que se dirige el tratamiento de los datos, en este caso, la obligación consistiría en el análisis de la solvencia determinado por la DCI y DCC (Deuring, 2022).

En este sentido, y con respecto al principio de minimización de datos mencionado en el capítulo segundo, los datos personales sólo podrán ser almacenados y tratados en la medida necesaria para su finalidad. Por ello, se tendrá que analizar si la finalidad pretendida se puede obtener con datos anonimizados. Si esto es posible, recurrir a datos identificables supondría una violación del principio de minimización de datos, por lo que únicamente recurriremos a la seudonimización cuando no haya otra opción. Para mejorar estos procesos de anonimización y seudonimización y cumplir con el principio de minimización de datos del artículo 5 RGPD (Campos, 2024), se están desarrollando herramientas como puede ser CIB PoP, una IA capaz de detectar automáticamente datos personales y todo contenido relevante para el RGPD para después eliminarlo o convertirlo en seudónimos (Deuring, 2022).

Asimismo, también deberá hacerse un esfuerzo por promover métodos de IA que utilicen menos datos. Entre estas técnicas pueden destacarse los modelos de aprendizaje federado (*Federated Learning*), donde los datos se mantienen en una serie de servidores sin ser transferidos, y el aprendizaje por transferencia (*Transfer Learning*), donde se aplican conocimientos aprendidos de un problema a otro, lo cual requeriría menos datos en general. Así estos modelos no solo ayudarán de cara al cumplimiento de los principios del RGPD sino que también permitirán que se preserve la privacidad en mayor medida (Khalid, 2023).

3.2. Protocolos y normativas de seguridad internos y externos

Otra medida clave consiste en el desarrollo de protocolos y normativas internas y externas para el uso de IA, así como la implementación de controles de seguridad y evaluación continua.

A nivel interno, las organizaciones deben desarrollar directrices en las que se establezcan qué datos se pueden compartir con las herramientas de IA y bajo qué condiciones. Estos protocolos deberán estar en constante supervisión y controlarse

regularmente a medida que avance el empleo de la IA y surjan nuevos casos de uso (Cloudfare, 2025). En concreto, en relación con el tratamiento de datos alternativos, será fundamental examinar las relaciones entre los atributos presentes en estos datos y las características sensibles prohibidas para que, en caso de darse una fuerte correlación entre una variable y una característica protegida, pueda considerarse un *proxy* y ser eliminada (Griffith, 2023).

A nivel externo, debería hacerse un esfuerzo por abordar los desafíos que genera la actual legislación de protección de datos, planteándose la corrección de determinados problemas por la vía jurídica. Lo cierto es que el actual RGPD no hace mención específica a la IA, pudiendo ser una solución la creación de un marco legislativo específico para la IA en materia de protección de datos que pueda reforzar y complementar el RGPD, permitiendo que se adapten aquellas disposiciones que no terminan de casar con las características y naturaleza de la IA (Barrat, 2023).

3.3. Tecnologías avanzadas de protección de datos

Dicho esto, y en relación con la seguridad de la IA, las soluciones de protección de datos tradicionales a veces no son lo suficientemente flexibles para abordar los desafíos que presenta la IA, cobrando gran relevancia las tecnologías avanzadas en protección de datos.

Una de las principales recomendaciones para mitigar estos riesgos es la implementación de soluciones de protección de datos que han sido específicamente diseñadas para anticipar amenazas relacionadas con IA. Estas soluciones son capaces de proteger el código y los datos confidenciales en cada uno de los entornos en los que se almacenan y, al mismo tiempo, adaptarse a las diferentes necesidades de seguridad y privacidad de cada organización. Empresas como Cloudflare han desarrollado plataformas avanzadas para minimizar los riesgos asociados a la IA, combinando protección frente a la pérdida de datos (DLP), esto es, sistemas que evitan la filtración accidental de datos confidenciales, con estrategias de seguridad en tiempo real, permitiendo que las entidades protejan sus datos con mayor eficacia y sean capaces de afrontar los desafíos relacionados con la seguridad de la IA (Cloudfare, 2025).

CAPÍTULO IV. CONCLUSIONES

Una vez analizadas todas estas cuestiones, se puede concluir que el empleo de la IA en la concesión de préstamos bancarios representa un avance significativo, sobre todo si elementos como el *machine learning* o el uso de datos alternativos se introducen en la ecuación. Gracias a estas técnicas, entre otras ventajas, se mejoran los instrumentos de análisis de solvencia, se reduce la morosidad y el fraude, se aumenta la eficiencia, se mejora la experiencia del usuario y se permite el acceso al crédito por aquellos grupos que, por la vía tradicional, hubieran quedado desproveídos de financiación.

No obstante, el uso de la IA también tiene sus inconvenientes. Aquellas características que hacen de la IA una herramienta de gran utilidad también originan un gran desafío que pone en riesgo los principios éticos de autonomía, prevención de daño, justicia y explicabilidad por los que debe guiarse la IA. Los sesgos algorítmicos, la falta de explicabilidad y los problemas con el tratamiento de datos son realidades ineludibles a las que las entidades financieras deben enfrentarse si desean hacer uso de modelos de IA para la concesión de préstamos bancarios.

En cuanto al sesgo algorítmico, los principales problemas identificados son la existencia de sesgos en los datos (mala calidad y asimetría de datos), así como en los procesos de la IA (sesgo de retroalimentación y de *proxy*). Para mitigar estos riesgos, es necesario implementar técnicas de tratamiento previo al entrenamiento de la IA (rebalanceo y remuestreo de datos, generación de datos sintéticos), tratamiento durante el entrenamiento (regularización y métodos adversariales) y tratamiento posterior al entrenamiento (ajuste de resultados), así como de supervisión activa y constante por parte de profesionales *(human in and over the loop)*.

Respecto a la explicabilidad, los problemas principales radican en la opacidad inherente de ciertos algoritmos de *machine learning* y la dificultad para comprender y auditar las decisiones automatizadas. Para abordar esto, se propone la adopción de sistemas híbridos que incorporen técnicas de inteligencia artificial explicable (XAI) como LIME, SHAP y FI, así como la restricción de la monotoneidad como técnica para mejorar la interpretabilidad de los modelos.

En relación con la protección de datos, los problemas clave son el conflicto con los principios de minimización de datos y limitación de finalidad del RGPD, así como los riesgos derivados del uso de datos personales alternativos y las potenciales brechas de seguridad y ciberataques. Para mitigar estos riesgos se recomienda la adopción de técnicas de anonimización y seudonimización de datos, así como protocolos internos complementados por regulaciones externas específicas para el uso de IA, así como el uso de tecnologías especializadas en fortalecer la seguridad informática de la IA.

En definitiva, este trabajo concluye que la IA tiene el potencial de transformar positivamente la concesión de préstamos bancarios, pero su implementación debe ser responsable y, únicamente a través de mecanismos de supervisión adecuados y la incorporación de enfoques explicables, justos y seguros, podrá desarrollarse la IA en su máximo exponente sin comprometer los principios éticos que la configuran.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Blanca López-Tello Martínez, estudiante de 5º E-3B de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "El uso de la inteligencia artificial en la concesión de préstamos bancarios: problemas éticos", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Crítico:** Para encontrar contra-argumentos a una tesis específica que pretendo defender.
3. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
4. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
5. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.
6. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
7. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 24 de marzo de 2025

Firma: Blanca López-Tello Martínez

BIBLIOGRAFÍA

Agencia Española de Protección de Datos. (2024). *Principios de Protección de Datos.* AEPD. https://www.aepd.es/derechos-y-deberes/cumple-tus-deberes/principios

Aldasoro, I., Doerr, S., Gambacorta, L., & Rees, D. (2024). *The impact of artificial intelligence on output and inflation* (BIS Working Papers No. 1179). Bank for International Settlements. https://www.bis.org/publ/work1179.pdf

Alonso-Robisco, A., & Carbó, J. M. (2023). Aprendizaje automático en modelos de concesión de crédito: oportunidades y riesgos. En S. Carbó, J. J. Ganuza, D. Peña, & P. Poncela (Eds.), *Análisis financiero y big data* (pp. 79–104). Funcas. https://www.funcas.es/wp-content/uploads/2023/05/Analisis-financiero-y-big-data_Capitulo-III.pdf

Aragüez Valenzuela, L. (2024). *Hacia la eticidad algorítmica en las relaciones laborales*. Laborum. https://dialnet.unirioja.es/descarga/libro/979447.pdf

Asociación Española de Banca. (2019). *Cómo mejorar la experiencia del usuario de banca con la inteligencia artificial*. AEB. https://www.aebanca.es/noticias/como-mejorar-la-experiencia-del-usuario-de-banca-con-la-inteligencia-artificial/

Ayerbe, A. (2020). *La ciberseguridad y su relación con la inteligencia artificial. Análisis del Real Instituto Elcano (ARI), (128).* https://media.realinstitutoelcano.org/wp-content/uploads/2021/10/ari128-2020-ayerbe-ciberseguridad-y-su-relacion-con-inteligencia-artificial.pdf

Aziz, S., & Dowling, M. (2018). Machine Learning and AI for Risk Management. En *Palgrave studies in digital business & enabling technologies* (pp. 33-50). https://doi.org/10.1007/978-3-030-02330-0_3

Bahoo, S., Cucculelli, M., Goga, X., & Mondolo, J. (2024). Artificial intelligence in finance: A comprehensive review through bibliometric and content analysis. *Financial Innovation, 10*(1), 123. https://www.researchgate.net/publication/377563906_Artificial_intelligence_in_Finance_a_comprehensive_review_through_bibliometric_and_content_analysis

Banco de España. (2022). *Documento de conclusiones sobre el desarrollo y los resultados del proyecto piloto "NDT - IA explicable en la gestión de Riesgos"*. https://www.bde.es/f/webbde/INF/MenuHorizontal/Servicios/Sandbox/Conclusiones_NeuroDecision_Technology.pdf

Banco de España. (2024). La inteligencia artificial en el sistema financiero. *Revista de Estabilidad Financiera*, 47, 15–30. https://www.bde.es/f/webbe/GAP/Secciones/Publicaciones/InformesBoletinesRevistas/RevistaEstabilidadFinanciera/24/1_REF47_Artificial.pdf

Barrat, L. P. U. (2023). *Data Protection in AI-Driven Systems: Understanding the EU's Legal and Regulatory Response Through the General Data Protection Regulation (GDPR)*. IE University Law School. https://www.researchgate.net/publication/385089185_Data_Protection_in_AI-Driven_Systems_Understanding_the_EU's_Legal_and_Regulatory_Response_Through_the_General_Data_Protection_Regulation_GDPR

Bishop, Christopher M. (2006) *Patrones de reconocimiento y aprendizaje automático.* Springer. https://github.com/peteflorence/MachineLearning6.867/blob/master/Bishop/Bishop%20-%20Pattern%20Recognition%20and%20Machine%20Learning.pdf

Breiman, L. (2001). *Random forests*. *Machine Learning*, *45*(1), 5–32. https://link.springer.com/article/10.1023/A:1010933404324

Campos Rivera, G. (2024). Credit Scoring como tratamiento de datos personales a la luz del RGPD. Análisis de su finalidad e influencia en los posibles usos secundarios

de los datos. *Revista de Derecho de la UNED (RDUNED)*, *33*, 111-150. https://revistas.uned.es/index.php/RDUNED/article/view/41926

Carpi Martín, R. (2023). Evaluación de solvencia y calificación crediticia en el reglamento europeo sobre Inteligencia Artificial (Ley de Inteligencia Artificial). En I. Herbosa Martínez & D. Fernández de Retana Gorostiza (Eds.), *Derecho e Inteligencia Artificial: el jurista ante los retos de la era digital* (pp. 71-112). Thomson Reuters (Azanzadi) https://merit.url.edu/ws/portalfiles/portal/38360638/Ficheros_de_solvencia_e_IA_.pdf

Carrascosa, A. (2024). ¿Cómo está revolucionando la IA el sector bancario y la concesión de créditos?. *Do better by Esade.* https://dobetter.esade.edu/es/IA-bancos-credito

Cloudflare. (2025). *Data Protection & AI.* https://www.cloudflare.com/es-es/the-net/data-protection-ai/

Coekelbergh, M. (2020). I*A Ethics/Ética de la inteligencia artificial*. Ediciones Cátedra. https://www.marcialpons.es/media/pdf/9788437642123.pdf

Collado-Rodríguez, J. (2023). La evaluación de la solvencia mediante el uso de sistemas de IA. *Revista CESCO de Derecho de Consumo*, 46, 41-67. https://dialnet.unirioja.es/descarga/articulo/8990540.pdf

Colmenarejo Fernández, R., (2018). Ética aplicada a la gestión de datos masivos. *Anales de la Cátedra Francisco Suárez. 52*, 113–129. https://revistaseug.ugr.es/index.php/acfs/article/download/6553/5674/16236

Comisión Europea. (2019). *Ethics guidelines for trustworthy AI*. High-Level Expert Group on AI. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

Comisión Europea. (2021). *Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión.* https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX%3A52021PC0206

Consejo de la Unión Europea. (2025). *Políticas sobre inteligencia artificial.* https://www.consilium.europa.eu/es/policies/ai-explained/

Cortina Orts, A. (2019). Ética de la inteligencia artificial. *Anales de la Real Academia de Ciencias Morales y Políticas, 96*, 379-394. https://www.boe.es/biblioteca_juridica/anuarios_derecho/abrir_pdf.php?id=ANU-M-2019-10037900394

De Lucas, L. (2024). Cómo la IA está mejorando la experiencia del cliente. *Zona Movilidad.* https://www.zonamovilidad.es/como-ia-esta-mejorando-experiencia-cliente

Deuring, F. (2022). Seudonimización y Anonimización: Protección de Datos con IA. *CIB.* https://www.cib.de/es/seudonimizacion-y-anonimizacion-proteccion-de-datos-con-ia/

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv (Cornell University)*. https://arxiv.org/abs/1702.08608

European Commission, High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI.* European Commission. https://www.ccdcoe.org/uploads/2019/06/EC-190408-AI-HLEG-Guidelines.pdf

Federación Bancaria Europea. (2021). *Artificial intelligence in the banking sector: Opportunities and challenges*. EBF_037419 - Artificial Intelligence in the banking sector - FINAL - for approval - CLEAN.docx

Felzmann, H., Fosch-Villaronga, E., Lutz, Ch., Tamo-Larrieux, A., (2020) Towards transparency by design for Artificial Intelligence. *Science and Engineering Ethics*. *26*(6), 3333–3361. https://link.springer.com/article/10.1007/s11948-020-00276-4

Fernández, A. (2019). Inteligencia artificial en los servicios financieros. *Boletín Económico*, 2, 1–10. https://repositorio.bde.es/handle/123456789/8448

Funcas. (2024). *La explicabilidad de la inteligencia artificial en la banca*. https://www.funcas.es/odf/la-explicabilidad-de-la-inteligencia-artificial-en-la-banca/?utm_source=chatgpt.com

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77(1), 5-47. https://ideas.repec.org/a/bla/jfinan/v77y2022i1p5-47.html

Giralt García, V. F. (2024). Prólogo al libro de Lucía Aragüez Valenzuela, *Hacia la eticidad algorítmica en las relaciones laborales.* Laborum. https://dialnet.unirioja.es/descarga/libro/979447.pdf

Gmyrek, P., Lutz, C., & Newlands, G. (2024). A Technological Construction of Society: Comparing GPT-4 and Human Respondents for Occupational Evaluation in the UK. *International Labour Organization.* https://www.researchgate.net/publication/377526649_A_Technological_Construction_of_Society_Comparing_GPT-4_and_Human_Respondents_for_Occupational_Evaluation_in_the_UK

González-Sendino, R., Serrano, E., & Bajo, J. (2024). Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. *Future Generation Computer Systems*, *155*, 384-401.https://www.researchgate.net/publication/378473107_Mitigating_bias_in_artificial_intelligence_Fair_data_generation_via_causal_models_for_transparent_and_explainable_decision-making

Goñi Sein, J. L. (2019). Innovaciones tecnológicas, inteligencia artificial y derechos humanos en el trabajo. *Documentación Laboral*, 117, 58-71. https://www.aedtss.com/wp-content/uploads/dl/N117/07%20Innovaciones%20tecnol%C3%B3gicas,%20inteligencia%20artificial%20y%20derechos%20humanos%20en%20el%20trabajo%20(Go%C3%B1i%20Sein).pdf

Griffith, M. A. (2003). AI lending and the ECOA: Avoiding accidental discrimination. *North Carolina Banking Institute Journal*, 27, 349-372. https://heinonline.org/HOL/Page?handle=hein.journals/ncbj27&id=372&collection=journals&index=

Heaven, W. D. (2021). Bias isn't the only problem with credit scores—and no, AI can't help. *MIT Technology Review*. https://www.technologyreview.com/2021/06/17/1026519/racial-bias-noisy-data-credit-scores-mortgage-loans-fairness-machine-learning/

Hoffman, R. R., Mueller, S. T., Klein, G. y Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv (Cornell University)*. https://arxiv.org/abs/1812.04608

Huang, Y., Qiu, H., Sun, T., & Wang, X. (2020). Credit risk evaluation in FinTech for SMEs: Evidence from China. *Journal of Banking and Finance*, *102*, 207-217. https://www.imf.org/en/Publications/WP/Issues/2020/09/25/Fintech-Credit-Risk-Assessment-for-SMEs-Evidence-from-China-49742

Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. *Yale JL & Tech.*, 18, 148-216. https://openyls.law.yale.edu/bitstream/handle/20.500.13051/7808/Hurley_Mikella.pdf

Iturmendi Rubia, J. M. (2023). La discriminación algorítmica y su impacto en la dignidad de la persona y los derechos humanos. Especial referencia a los inmigrantes. *Derechos Humanos y Religión,* *18*(2), 1–29. https://doi.org/10.18543/djhr.2910

Joseph, Andreas. 2020. *Parametric inference with universal function approximators* (Bank of England Working Papers No. 784). Bank of England. https://ideas.repec.org/p/boe/boeewp/0784.html

Karjka, S. (2024). Transforming banking: The power of chatbots in customer service. *Boost.ai*. https://boost.ai/blog/banking-chatbot/

Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in Biology and Medicine, 158*, 106848. https://doi.org/10.1016/j.compbiomed.2023.106848

Kou, G., Chao, X., Peng, Y., Alsaadi, F. E., & Herrera-Viedma, E. (2019). Machine Learning Methods for Systemic Risk Analysis in Financial Sectors. *Technological and Economic Development of Economy, 25,* 716-742. https://www.researchgate.net/publication/333470596_MACHINE_LEARNING_METHODS_FOR_SYSTEMIC_RISK_ANALYSIS_IN_FINANCIAL_SECTORS

López Martínez, F., & García Peña, J. H. (2024). IA y sesgos: una visión alternativa expresada desde la ética y el derecho. *Informática y Derecho. Revista Iberoamericana de Derecho Informático (2.ª época), 1*(15), 109-121. *https://revistas.fcu.edu.uy/index.php/informaticayderecho/article/view/4738*

Lousada Arochena, J. F. (2024). *I*nteligencia artificial y sesgos discriminatorios: ¿Es necesario un nuevo concepto de discriminación algorítmica? *IgualdadES, 11,* 97-123. https://doi.org/10.18042/cepc/IgdES.11.04

Lundberg, S. M., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. *arXiv (Cornell University)*. https://arxiv.org/abs/1705.07874

McKinsey & Company. (2024). *Generative AI in banking: The next frontier for risk and compliance*. *McKinsey & Company*. https://www.mckinsey.com

Megías Quirós, J. J. (2022). Derechos humanos e inteligencia artificial. *Dikaiosyne: Revista semestral de filosofía práctica,* 37, 140–163. https://rodin.uca.es/bitstream/handle/10498/31390/10%20Dikaiosyne%20Dhs%20e%20IA.pdf?sequence=1&isAllowed=y

Mitrou, L. (2018). *The General Data Protection Regulation: A law for the Digital Age?* In T. Synodinou et al. (Eds.), *EU Internet Law, Regulation and Enforcement* (pp. 19-57). Springer. https://www.researchgate.net/publication/320984809_The_General_Data_Protection_Regulation_A_Law_for_the_Digital_Age

Nallakaruppan, M. K., Chaturvedi, H., Grover, V., Balusamy, B., Jaraut, P., Bahadur, J., Meena, V. P., & Hameed, I. A. (2024). Credit Risk Assessment and Financial Decision Support Using Explainable Artificial Intelligence. *Risks*, *12*(10), 164. https://www.researchgate.net/publication/385058079_Credit_Risk_Assessment_and_Financial_Decision_Support_Using_Explainable_Artificial_Intelligence

Narang, A., Vashisht, P., & Bajaj, S. B. (2024). Artificial Intelligence in Banking and Finance. *International Journal Of Innovative Research In Computer Science & Technology*, *12*(2), 130-134. https://www.researchgate.net/publication/380119377_Artificial_Intelligence_in_Banking_and_Finance

Nguyen, A., & Martínez, M. R. (2019). MonoNet: Towards Interpretable Models by Learning Monotonic Features. *arXiv (Cornell University)*. https://arxiv.org/pdf/1909.13611

O'Leary, D. E. (2013). Artificial Intelligence and Big Data. *IEEE Intelligent Systems*, *28*(2), 96-99. https://doi.org/10.1109/MIS.2013.39

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, *2*, 13. https://pubmed.ncbi.nlm.nih.gov/33693336/

Palma Ortigosa, A. (2021). *Régimen jurídico de la toma de decisiones automatizadas y el uso de sistemas de inteligencia artificial en el marco del derecho a la protección de datos personales* (Tesis doctoral). Universitat de València. https://www.uv.es/cotino/INDICE%20CON%20PORTADA.pdf

Parlamento Europeo. (2020). *¿Qué es la inteligencia artificial y cómo se usa?* https://www.europarl.europa.eu/topics/es/article/20200827STO85804/que-es-la-inteligencia-artificial-y-como-se-usa

Parlamento Europeo y Consejo de la Unión Europea. (2021). *Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión*. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:52021PC0206

Rebolledo Vivas, R. (2024). *Revolución en la gestión del riesgo de crédito con inteligencia artificial*. https://es.linkedin.com/pulse/revolución-en-la-gestión-del-riesgo-de-crédito-con--ha5ee

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *arXiv (Cornell University)*. https://arxiv.org/abs/1602.04938

Ridzuan, N., Masri, M., Anshari, M., Fitriyani, N. L., & Syafrudin, M. (2024). AI in the financial sector: The line between innovation, regulation and ethical responsibility. *Information*, *15*(8), 432. https://www.mdpi.com/2078-2489/15/8/432

Rivas Vallejo, P. (2020). *La aplicación de la inteligencia artificial al trabajo y su impacto discriminatorio*. Aranzadi. https://tienda.aranzadilaley.es/la-aplicacion-de-la-inteligencia-artificial-al-trabajo-y-su-impacto-discriminatorioduo?srsltid=AfmBOorfEa18z8mLfMjkJEug2kZFZ1Ub_kyILmXbB5UQOj5t_u9oHq0a

Rizinski, M., Peshov, H., Mishev, K., Chitkushev, L. T., Vodenska, I., & Trajanov, D. (2022). Ethically responsible machine learning in Fintech. *IEEE Access, 10*, 97531-97554. https://www.researchgate.net/publication/363104797_Ethically_Responsible_Machine_Learning_in_Fintech

Rodríguez Fernández, M. L. (2024). Inteligencia artificial, género y trabajo. *Temas Laborales*, *171*, 11-39. https://dialnet.unirioja.es/descarga/articulo/9539778.pdf

Ruiz, F. J. B. (2022). Paradoja de la transparencia en la IA. *Revista Internacional de Pensamiento Político, 17, 2*61-272. https://doi.org/10.46661/revintpensampolit.7526

Sadok, H., Sakka, F., & Maknouzi, M. E. H. E. (2022). Artificial intelligence and bank credit analysis: A review. *Cogent Economics & Finance*, *10*(1). https://www.tandfonline.com/doi/full/10.1080/23322039.2021.2023262#abstract

Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V., & Napolitano, A. (2008). Resampling or reweighting: A comparison of boosting implementations. En *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, 445–451. https://www.researchgate.net/publication/221416853_Resampling_or_Reweighting_A_Comparison_of_Boosting_Implementations

Stahl, B. C., Antoniou, J., Ryan, M., Macnish, K., & Jiya, T. (2022). Organisational responses to the ethical issues of artificial intelligence. *AI & Society 37*(1), 23-37. https://link.springer.com/article/10.1007/s00146-021-01148-6

Stevens, A., Deruyck, P., Van Veldhoven, Z., & Vanthienen, J. (2020). Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva. *2021 IEEE Symposium Series On Computational Intelligence (SSCI)*, 1241-1248. https://doi.org/10.1109/SSCI47803.2020.9308371

Stoyanovich, J., Howe, B., & Jagadish, H. V. (2020). *Responsible Data Management*. Proceedings of the VLDB Endowment, *13*(12), 3474-3488. https://www.vldb.org/pvldb/vol13/p3474-asudeh.pdf

Todolí Signes, A. (2024). Algoritmos productivos y extractivos. Cómo regular la digitalización para mejorar el empleo e incentivar la innovación, Navarra, España, Aranzadi, 2023. *Revista de la Facultad de Derecho de México*, *74*(290), 475-480. https://www.researchgate.net/publication/388001871_Todoli_Signes_Adrian_Algoritmos_productivos_y_extractivos_Como_regular_la_digitalizacion_para_mejorar_el_empleo_e_incentivar_la_innovacion_Navarra_Espana_Aranzadi_2023

Tribunal de Justicia de la Unión Europea. (2023). *Sentencia del Tribunal de Justicia (Sala Primera) de 7 de diciembre de 2023. OQ contra Land Hessen*. https://op.europa.eu/es/publication-detail/-/publication/acd82fb9-94e3-11ee-b164-01aa75ed71a1/language-es

Vlaicu, R. (2024). Can AI Technologies Help Expand Credit Access?. *Ideas Matter*. https://blogs.iadb.org/ideas-matter/en/can-ai-technologies-help-expand-credit-access/

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76-99. https://academic.oup.com/idpl/article-abstract/7/2/76/3860948?redirectedFrom=fulltext

Xi, B. (2021). Adversarial Machine Learning for Cybersecurity and Computer Vision: Current Developments and Challenges. *arXiv (Cornell University)*. https://arxiv.org/abs/2107.02894

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *arXiv (Cornell University)*. https://arxiv.org/pdf/1907.00503

Zest AI. (2019). *There's a fix to the problem of biased algorithms in lending.* https://www.zest.ai/learn/blog/theres-a-fix-to-the-problem-of-biased-algorithms-in-lending/

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). *Mitigating unwanted biases with adversarial learning*. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* 335–340. Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3278721.3278779